Dissertation

Medical Information Search in Semi-Structured Data

submitted by

Markus Eduard KREUZTHALER
MSc, BSc

for the Academic Degree of

Doctor of Medical Science
(Dr. scient. med.)
at the
Medical University of Graz

Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz
Head: Univ.-Prof. Dipl.-Ing. Dr. techn. Andrea Berghold

---

under the Supervision of

Univ.-Prof. Dr.med. Stefan Schulz
Univ.-Prof. Dipl.-Ing. Dr. techn. Andrea Berghold
Prof. Dr. habil. Henning Müller

Graz, October 2015

Dissertation Committee

Univ.-Prof. Dr.med. Stefan Schulz
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz
Auenbruggerplatz 2
8036 Graz

Univ.-Prof. Dipl.-Ing. Dr. techn. Andrea Berghold
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz
Auenbruggerplatz 2
8036 Graz

Prof. Dr. habil. Henning Müller
University of Applied Sciences Western Switzerland, Sierre (HES-SO)
TechnoArk 3
3960 Sierre, Switzerland

Declaration

I hereby declare that this thesis is my own original work and that I have fully acknowledged by name all of those individuals and organisations that have contributed to the research for this thesis. Due acknowledgement has been made in the text to all other material used. Throughout this thesis and in all related publications I followed the guidelines of "Good Scientific Practice".

Eidesstattliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbständig angefertigt und abgefasst, und jene Personen und Institutionen, die am Zustandekommen der Forschungsdaten beteiligt waren, namentlich genannt habe. Andere als die angegebenen Quellen habe ich nicht verwendet und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen habe ich als solche kenntlich gemacht. Die Arbeit an der Dissertation und daraus entstandener Publikationen wurde gemäß den Regeln der „Good Scientific Practice" durchgeführt.

Graz, 28th October 2015 Markus Kreuzthaler

Danksagung

Diese Arbeit wurde im Jahr 2015 am Institut für Medizinische Informatik, Statistik und Dokumentation an der Medizinischen Universität Graz verfasst. Teile der Arbeit wurden durch das Toolset der Firma Averbis GmbH unterstützt.

Ich möchte mich hierbei vor allem bei alle beteiligten Personen des Instituts bedanken, die mich bei verschiedensten Aspekten der Arbeit unterstützt haben (Infrastruktur, Inhalt, Feedback). Ich habe in den unteschiedlichen Bereichen der Medizinfinformatik über die Jahre vieles mitgenommen und mich dennoch auf ein Gebiet spezialisieren können. Des weiteren gilt ein großer Dank Univ.-Prof. Dr.med. Stefan Schulz und Univ.-Prof. Dipl.-Ing. Dr. techn. Andrea Berghold die es mir ermöglicht haben mich ausführlich mit dem Inhalt dieser Dissertation zu beschäftigen und mir über die Jahre einen sicheren Rückhalt gegeben haben. Auch der gesamten BST Projektgruppe möchte ich einen großen Dank für die angenehme Arbeitsatmosphäre und das gute Gesprächsklima aussprechen. And thank you Marcus for always having an open ear.

Ein weiterer Dank gilt meiner Familie, meinen Freunden und besonders Carine, die mich während meiner intensiven Phase der Arbeit begleitet hat.

> "*Coming back to where you started is not the same as never leaving.*"
> A Hat Full of Sky, Terry Pratchett (1948-2015).

Graz, am 28. Oktober 2015 Markus Kreuzthaler

Abstract

Computer systems for clinical information management store large amounts of textual data in medical records. Finding reports and patient summaries use semi-structured document templates. Coded data, using controlled vocabularies, are mainly restricted to accounting, research, and quality assurance. Unstructured or semi-structured content is difficult to analyse, although there are multiple use cases for content retrieval from clinical texts which would benefit from semantically enhanced retrieval functionalities.

This thesis focuses on the investigation of clinical narratives in combination with improved semantic indexing and extraction systems for patient-based decision making. It addresses the development and evaluation of technical solutions to support health professionals and researchers in retrieving targeted patient-related information in a timely and efficient way, according to their information needs. The information to be searched for is constituted within medical free text from various clinical domains in a hospital environment. Different state of the art approaches are explored to what extent they can be adapted to the domain and how they can be optimized to apply to clinical professionals' information needs.

These approaches, applied to anonymized clinical textual data, show the potential of adapted solutions to medical domains and related sublanguages for enhanced information retrieval. The following search scenarios have been investigated: collection-based patient search and patient-based document search. Within these search scenarios, enhanced text processing methods have shown their applicability to support domain expert retrieval. The results show the trade-off of clinical information systems and the possibilities of novel frameworks and technologies for unstructured information processing. For selected clinical information system content, they can bridge the gap between *patient-based storage systems* and *disease-related search systems*.

Zusammenfassung

Klinische Informationssysteme enthalten große Mengen von textuellen Daten für patientenbasierte medizinische Aufzeichnungen, wobei diese meist in semi-strukturierten Dokumentvorlagen eingebettet sind. Die Verwendung von kodierten Daten und einem kontrollierten Vokabular, ist meist auf die Abrechnung medizinischer Leistungen, Forschung und Qualitätssicherung beschränkt. Unstrukturierte oder semi-strukturierte Informationen sind schwierig zu analysieren, obwohl es jedoch Anwendungsfälle der klinischen Dokumentenrecherche gibt, die von einer verbesserten Semantik profitieren würden.

Diese Arbeit konzentriert sich auf die Entwicklung und Bewertung technischer Lösungen für Gesundheitsfachkräfte und Forscher bei der Suche nach patientenbezogenen Informationen und wie diese dabei effizient, je nach Informationsbedarf, unterstützt werden können. Die Datenbestände, die dabei untersucht werden, sind aus verschiedenen klinischen Abteilungen entstanden und wurden innerhalb eines klinischen Informationssystems abgespeichert. Verschiedene dem aktuellen Stand der Technik existierende Frameworks, Methoden und Lösungen aus Industrie und Forschung werden untersucht, in wie weit sie sich auf die jeweilige Sprachendomäne anpassen und optimieren lassen, um gängige Suchanfragen von medizinischen Experten zu unterstützen.

Die verschiedenen Ansätze wurden auf anonymisierte klinische Texte angewandt und zeigen das Potential angepasster Suchlösungen für die jeweilige Sprachendomäne für eine bessere Informationsrecherche in der Medizin. Folgende Rechercheszenarien wurden dabei untersucht: dokumentenbasierte Patientensuche und patientenbasierte Dokumentensuche. In diesen Anwendungsfällen können erweiterte Textverarbeitungsmethoden den medizinischen Domänenexperten in seiner Arbeit unterstützen. Die Ergebnisse zeigen die Möglichkeiten und Limitationen neuer Frameworks und Technologien für die Verarbeitung von unstrukturierter Information in klinischen Informationssystemen auf. Für ausgewählte Inhalte in klinischen Informationssystemen, können sie eine Verbindung zwischen *patientenbasierten Speichersystemen* und *krankheitsorientierten Suchsystemen* darstellen.

Contents

Glossary

AP	Average Precision
AUC	Area Under the Curve
CIS	Clinical Information System
CLEF	Clinical E-Science Framework
DFA	Deterministic Finite Automaton
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders - Revision Four
EHR	Electronic Health Record
ELGA	Austrian Personal Health Record
ETL	Extract Transform Load
HIPAA	Health Insurance Portability and Accountability Act
HMM	Hidden Markov Model
ICD	International Statistical Classification of Diseases and Related Health Problems
ICD-10	International Statistical Classification of Diseases and Related Health Problems - Revision Ten
ICD-9	International Statistical Classification of Diseases and Related Health Problems - Revision Nine
ICD-9-CM	International Statistical Classification of Diseases and Related Health Problems - Revision Nine - Clinical Modification
ICD-O-3	International Classification of Diseases for Oncology - Revision Three
IE	Information Extraction
IMI	Institute for Medical Informatics, Statistics and Documentation
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
LIS	Laboratory Information System
LSA	Latent Semantic Analysis

LSI	Latent Semantic Indexing
MAP	Mean Average Precision
MCGE	Medical Clinic Gastro-Enterology
MedLEE	Medical Language Extraction and Encoding System
MeSH	Medical Subject Headings
NDCG	Normalized Discounted Cumulative Gain
NER	Named Entity Recognition
NFA	Nondeterministic Finite Automaton
NLP	Natural Language Processing
PLSA	Probabilistic Latent Semantic Analysis
POS	Part of Speech
ROC	Receiver Operator Characteristic
SDS	Study Documentation System
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SQL	Structured Query Language
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TNM	Tumor Node Metastases
UIMA	Unstructured Information Management Architecture
UMLS	Unified Medical Language System
VSM	Vector Space Model

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Clinical Information Systems (CISs) comprise of a number of IT subsystems in healthcare institutions and their components largely differ according to the requirements in their underlying structure and content. The system architecture has to reflect the fact that a broad range of patient-related information is safely stored within heterogeneous data environments. Highly structured data environments used e.g. in Laboratory Information Systems (LISs) therefore coexist with large volumes of semi- or unstructured data sources like images or text. The amount of available free-text narratives in CISs, reflects the fact that major parts of patient-related information is only available in a semi-structured and non-standardized form.

The use of controlled vocabularies like the International Statistical Classification of Diseases and Related Health Problems - Revision Ten (ICD-10) system is generally restricted to encode basic information about main diseases and procedures. This complicates the retrieval of patient data when a variety of search criteria has to be combined. The retrieval or extraction of relevant information from semi- or unstructured data can be improved by enriching the textual content by semantic annotations. According to the clinical scenario, different natural language technologies may be used for processing these sources.

Natural language technologies receive more and more attention in CISs as narratives continue constituting the main carrier of information in most clinical disciplines, whereas reuse of this information beyond the primary intention, i.e. entire documents sequentially being read by humans, is increasingly valued. These so called secondary use scenarios (e.g. cohort building, quality management, decision support, patient safety) have directed the attention to apply natural language technologies to the analysis of unstructured content in Electronic Health Records (EHRs). There are different clinical scenarios where natural language technologies are expected to support the retrieval process according to specific information needs. Current technology allows the application of sophisticated text processing algorithms to millions of texts in a reasonable amount of time which also has to be considered in a clinical setting.

The right choice of the methods must equally consider peculiarities of medical language, which largely deviates from proof-read texts like textbooks or scientific literature. At the point of care, the most important requirement for a clinical text is that it carries information from the author to the reader. Both share the same terms and abbreviations, know the standard contexts, and are tolerant regarding the violation of grammar and spelling rules provided the text is clearly and quickly understandable. Given that such texts are produced under time constraints, the result is of such quality that fulfills the inter-professional communication needs, but is highly challenging when it comes to computer-based processing.

1.2 Objectives

The main objective of this work is to investigate how language technologies can be applied within a clinical environment and which performance can be achieved for the specific use case motivated within different clinical settings. These settings for enhanced Information Retrieval (IR) and Information Extraction (IE) will be presented in this thesis. The approaches will be introduced in the light of the state of the art as manifested in scientific literature and combined in a way which is beyond state of the art. Selected methods will be motivated, implemented and evaluated. State of the art frameworks and technologies for unstructured information management processing will be used, with methods chosen and customized for the specific problem domain. In contrast to the English speaking community where the application of language technologies to clinical narratives has a long tradition and their performance is constantly assessed in scientific challenges, German language lacks resources and corpora, especially regarding the biomedical discourse area. This hampers a combined research effort in this direction. The work presented in this thesis motivates the application of language technologies and solutions within diverse clinical settings. It ends with a discussion about how these technologies should be applied in future clinical information environments under the hypothesis of emerging technologies, with respect to IR and IE.

1.3 Outline

The work is structured as follows:

Chapter 2 provides background knowledge to the main aspects of this work. It introduces different levels of granularity of data structures and content, and a detailed insight into medical language in general, in the light of specific challenges to automated text processing. Different subareas of the field of human language technologies are outlined and contextualized regarding this work. Evaluation metrics are described, which proved useful to assess the performance of methods for automated language processing. The chapter is concluded by a survey of freely available data sets and scientific challenges, as well as by available clinical language technology systems.

Chapter 3 gives a detailed description of the core methods and models in use for four different language technology scenarios. In Chapter 4 the first scenario describes and evaluates

a clinical document classification use case. This is followed by a detailed investigation of a clinical IR use case in Chapter 5. Chapter 6 exposes a language technology scenario for clinical IE. The last of the four different use cases highlighted and evaluated in this work is presented in Chapter 7 providing a solution on a certain component within a clinical natural language pipeline.

The last Chapter 8 concludes this work and gives an outlook how results of this work could influence future CISs taking into account applied language technologies.

Chapter 2

Background

2.1 Semi-structured data

According to Holzinger et al. [1] a taxonomy of data is defined via two axes. The first one defines the level of structuredness, varying from highly structured to semi-structured data sources to unstructured data-representation forms. The second axis defines the level of standardization of the content. A controlled vocabulary like the Medical Subject Headings (MeSH) can be interpreted as having a high level of standardization (∼27,500 descriptors and ∼220,000 entry terms, manually revised and updated), together with a highly structured data environment defined via an alphabetical and a hierarchical structure. Most of the data from the CIS under investigation in this thesis were extracted and de-identified based on an Extract Transform Load (ETL) workflow using Talend Open Studio [2], resulting in a XML-based data structure which is categorized as semi-structured data environment within the level of standardization of clinical narrative content.

2.2 Medical language

Written language is used in a type of communication between humans. The complexity of the understanding of written language depends on the domain of the discourse as well as on characteristics of domain-specific sublanguages, which often heavily deviate from syntactic and stylistic conventions of standard languages. This is necessary for understanding the characteristics of biomedical texts, and the large differences that distinguish the language of scientific articles from the language of patient-related narratives, as the following text snippet demonstrates (with explanations in parentheses) [3]:

```
Ca.  2 x 1 cm, große ovaläre Verschattung im UF li.  Rez.  re.
lat.  frei, li.teilhärent
```

("An approximately 2 by 1 cm sized, oval-shaped opacity (in the chest X-ray) over (-laying) the left lower lobe; pleural recess (on the right side of the lobe) free of fluids, on the left side partly adhesive (after non-recent inflammation)"). As much as such highly

condensed text is understandable for physicians, it challenges any computer-based morphological and syntactic processing, as well as semantic interpretation. Typical language idiosyncrasies have to be considered for setting up an advanced Natural Language Processing (NLP) pipeline for clinical narratives: ambiguous terms, acronyms, abbreviations, single-word compounds, derivations, spelling variants, uncorrected spelling, typing and punctuation errors, jargon expressions, telegram style, non-standardized numeric expressions, and non-standard variations of negations. Patterson et al. [4] showed that different medical domains form distinct sublanguages of the overall expert jargon.

2.3 Language technologies

In this thesis, language technologies applied to clinical narratives to support medical information search in general are subdivided according to the use case and application scenario. The sub-areas that are addressed under the main topic of language technology scenarios are explained in the following sections.

2.3.1 Document classification

"The goal of text categorization is the classification of documents into a fixed number of predefined categories. Each document can be in multiple, exactly one, or no category at all." [5]

Basically, in this research area two different methodological approaches can be distinguished. On the one hand, knowledge engineering approaches use rules [6], often based on regular expressions to classify documents into a predefined set of categories. They heavily depend on the input of human experts who formulate decision rules. In the biomedical domain, the performance requirements are usually high, which results in a time intensive task.

In contrast, for supervised machine learning approaches documents are often represented as bag-of-words [7]. The underlying profound statistical models are trained on annotated corpora to build the classifier. It is argued that the knowledge engineering approach is shifted from the human to the machine in this case. To reach acceptable performance for the statistical classifier the efforts of feature engineering can be high and diminishes the advantages of this method. Furthermore, the resulting "black box" of the trained classifier is unintuitive as most of the time it does not explain classification results like in rule-based systems. Another line of document classification is clustering. Clustering approaches are unsupervised, i.e. they categorize documents without initial training annotations [8–10].

2.3.2 Information retrieval

"IR is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." [11]

Figure 2.1 illustrates a general information access/retrieval process, which starts with an information need. This information need can be resolved by choosing a computer system where the users intend to resolve their lack of information about a certain topic. In the next step, the information need is translated into the query syntax of the chosen retrieval system. The results of this transformation is, e.g., an Structured Query Language (SQL) expression for querying a database system, or a natural language search expression (usually a list of search terms) used as input for a Web retrieval system. In the following, the generated query is sent to the retrieval system and returns a retrieval result. This result is then assessed. If the information need could be resolved the process is finished, otherwise the search system query can be refined to gain better results. This generic process shows that the evaluation step within this model is fundamental. It is done, intuitively, by each user of an IR system. In systematic IR evaluation this is done experimentally against a gold standard. The requirements for IR gold standards are detailed in the following.

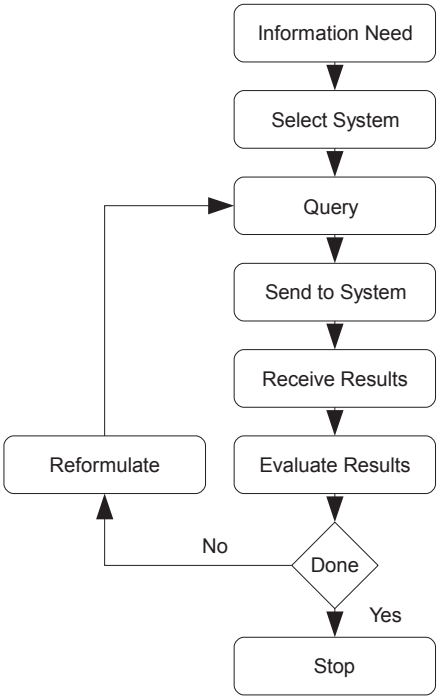


Figure 2.1: Standard model of the information access/retrieval process according to Baeza-Yates et al. [12].

Information retrieval evaluation

Expanding the definition of Manning et al. [11], Harter and Hert [13] (i) a representative sample size of a collection of documents within an operational setting of interest, (ii) a defined set of information needs ($\geq$25 [14]) and (iii) a relevance statement (relevant or non-relevant) for each document per information need is required for IR evaluation. This human respectively domain expert annotated document pool forms a so called gold

standard (ground truth judgment of relevance). For inter annotator agreement Cohen's kappa κ is used most of the time [15].

Having a gold standard at a certain quality level, metrics for *unranked* (e.g. precision, recall and F-measure [12]) or *ranked* (e.g. R-Precision, Precision at k, Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG) [11] or bpref [16]) retrieval results are used for IR performance evaluation. Besides these statistical evaluation measurements, six other levels according to Saracevic [17] should be taken into account: the engineering level, the input level, the processing level, the output level, the user and use level, and the social level. IR systems in general must also consider human factors like human information behavior to satisfy user needs beyond optimized retrieval results.

Collection-based patient search

In this work two main different retrieval scenarios are defined. The first one is collection-based patient search. Here, according to an information need, clinical documents are searched and patient IDs are returned according to the criteria that are applied to all or selected documents for each patient. This is typical for secondary use scenarios where patient cohorts are built, e.g., to support hypothesis generation. So far, CISs and their underlying patient centric architecture do not provide sufficient support for this kind of search, which makes this scenario complex and time consuming.

Patient-based document search

Patient-based document search defines the use case that within a patient's document collection, documents according to an information need are searched for. This use case is especially relevant where the amount of documents per patient is so high that browsing through documents becomes too time consuming. This scenario becomes more and more relevant in the context of chronic diseases, aging populations, personalized medicine, in parallel with the overall increase in specialized healthcare services. The need for this scenario is also driven by future eHealth platforms that aggregate data and documents from numerous providers, such as planned for the Austrian Personal Health Record (ELGA) platform, from which hundreds of documents can be accessed. Even in a single hospital, up to 80 discharge summaries per patient can be found [3]. Apart from searching within the document collection that belongs to one single person, the scenario can also be expanded to documents of multiple patients, addressing other use cases such as quality assurance or case-based medical education, among others.

2.3.3 Information extraction

"IE involves extracting predefined types of information from text. In contrast, IR is focused on finding documents and has some very popular examples such as the Google or PubMed search engines. IR returns documents whereas IE returns information or facts."

This explanation by Meystre et al. [18] of IE clearly delineates this field from IR. IE is often seen as sub-domain of NLP and has many application areas. Named Entity Recog-

nition (NER) is used to identify proper names in texts, such as person, institution or drug names. In a broader sense, NER also identifies controlled terms whatsoever, e.g. disease names. According to the Health Insurance Portability and Accountability Act (HIPAA) criteria [19], 18 types of so-called named entities (in a strict sense) have to be removed from a clinical narrative so it can be considered de-identified. In many cases, prior to IE, tasks are performed like text cleansing e.g. spelling correction or word sense disambiguation. IE methods often rely in their processing steps on prior annotations like Part of Speech (POS) tags or parser annotations, which represent the syntactical structure of a sentence in a tree based hierarchy. The extraction of contextual information is equally important and includes the following three parts according to Chapman et al. [20]: negations (affirmed, negated), temporality (recent, hypothetical, historical) and identifying the experiencer of a certain event. An important application for IE is the annotation of narratives with terminology codes, such as International Statistical Classification of Diseases and Related Health Problems (ICD) or Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT). This is often referred to as concept mapping and is an important step for enriching documents with coded metadata, which serve many use cases such as the generation of problem lists, support for administrative coding, adverse-event detection [21, 22], syndrome surveillance, and clinical decision support.

For all these areas the distinction between rule-based and statistical machine learning methods is relevant, both with their pros and cons mentioned in Section 2.3.1. Therefore, quality-assured gold standard annotations for training and testing of IE systems are of utmost importance for a reliable quality statement [23–26]. The automatically added annotations are often used in a second step in semantically enhanced IR scenarios, as well as in text mining scenarios that rely on information extracted from narratives, like stratifying disease-finding, drug-disorder or disease-disease relations via concept co-occurrence [27].

2.3.4 Natural language processing

"NLP is an automated technique that converts narrative documents into a coded form that is appropriate for computer-based analysis. Capabilities that NLP provides in the context of healthcare include parsing a sentence into its component structures, understanding the medical vocabulary and clinical terms used, disambiguating the context in order to interpret the clinical terms correctly within the broader context of the documentation, and representing the processed information for further use" [28].

Within a language technology scenario in this thesis, NLP is used to set up a proper processing pipeline for text annotations (e.g. tokenization, POS tagging, abbreviation detection, sentence splitting). Document classification and IE tasks can be seen as pipeline components within an aggregated NLP scenario. The distinction of the different subcategories in this work according to several language technologies is motivated according to different use case scenarios for processing clinical narratives within a hospital environment. Therefore, NLP is described at the same level as IE, IR and document classification in this section, because one experiment described in this thesis specifically focuses on one showcase aspect within a clinical NLP pipeline. Thus, a clear four-level language technology investigation of the different scenarios described above can be fulfilled.

NLP research has a long tradition within a large community (computational linguistics) and several well established conferences (e.g. ACL [29], EACL [30], SGIR [31], ISCA [32], EAMT [33]). In contrast, clinical NLP is a relatively small sub-area, especially when compared to bioNLP. One reason for this is the difficulty of accessing real data due to privacy concerns, which leads to a lack of shared data of gold standards. This is gradually changing within the English speaking community, as in the U.S. scientific challenges have been established to foster clinical NLP developments, principally the i2b2 challenges [34], SemEval [35, 36], and the TREC Medical Tracks [37] 2011 and 2012 with respect to processing clinical narratives. Furthermore, the International Workshop on Health Text Mining and Information Analysis [38] (Louhi) should be mentioned with a special focus on medical data in European languages.

The situation for the German language in this research area is not satisfying, due to the lack of a joined effort to foster research for processing clinical narratives in this language. The lack of openly available gold standards leads to research and development efforts that cannot be disclosed. Therefore the comparison with other new approaches is almost impossible. The JULIE Labs in Germany make their *language models* openly available so that other research institutions can test their pipeline components (e.g. sentence detection). They also made their Unstructured Information Management Architecture (UIMA) [39] based NLP framework openly accessible. Other research institutions that have considerable merits in clinical NLP are the Mayo Clinic [40] (cTakes), the Veterans Affairs network of hospitals [18, 41] (The Leo framework - The VINCI-developed NLP infrastructure using UIMA) and the Columbia-Presbyterian Medical Center [42, 43] (Medical Language Extraction and Encoding System (MedLEE)).

2.4 Data sets and frameworks

As mentioned before, clinical language technology systems are applied and tested in different challenges within academia. The following paragraph gives a rough overview of clinical free-text data sets that were used in these challenges. All of these data sets are available from English-speaking countries whereas no such open data sets exist in German despite the fact there are close to 100 million inhabitants [44], most of them having a footprint of clinical free text documentation in CISs. The following review is an extended version of Kreuzthaler et al. [45], based on Roberts et al. [46].

Within the Computational Medicine Challenge [47] 1,954 reports were assigned with International Statistical Classification of Diseases and Related Health Problems - Revision Nine - Clinical Modification (ICD-9-CM) codes from a radiology department. The collections initial purpose was to compare different classifiers [47]. ImageCLEFMed contains about 50,000 images for content-based image retrieval. The annotated textual information per picture can also be used for text-based IR evaluation [48, 49]. 60 clinical notes are tagged with functional disorders in the Ogren corpus [50]. The Clinical E-Science Framework (CLEF) corpus [51] contains 167 documents annotated with drugs, diseases, body regions etc. and relations between them. Within i2b2 there are several challenges. The initial corpus consists of 889 anonymized clinical summaries with annotations for evaluating de-identification. A subset of these summaries are tagged with information about

the patients' health status [52, 53]. One of the most recent challenges in 2014 was de-identification and cardiovascular risk factor identification over time. The BLULab NLP Repository was intended to be a resource for long term clinical NLP research, but it is no longer accessible. The University of Pittsburgh [54] managed access to a pool of 100,000 de-identified clinical narratives from various U.S. hospitals tagged with International Statistical Classification of Diseases and Related Health Problems - Revision Nine (ICD-9) codes during 2007. The TREC Medical Tracks [37] in 2011 and 2012 were started with the aim to advance research on providing content-based access to free-text fields in medical records. The BLULab NLP Repository was used as a resource in both cases. The recent SemEval 2014-2015 [55, 56] challenges contained tasks for abbreviation/acronym/entity detection, Unified Medical Language System (UMLS) concept mapping and timeline extraction applied to clinical narratives [57].

Within the German speaking community the following efforts have to be mentioned beside the fact that these data sets were not made openly available to the clinical NLP community: In the medSynDiKaTe project [58] 90 pathology reports on gastro-intestinal diseases were annotated and used in an advanced IE approach. The FReiburg Annotated MEDicine text corpus (FraMed) was POS tagged by Wermter and Hahn [59] and Faessler et al. [60]. The corpus consists of ∼6,500 sentences of different domains (discharge summaries, pathology reports, histology reports, surgery reports, textbook, consumer texts). Geierhofer and Holzinger [61] annotated ∼3,000 pathology reports for setting up a rule-based approach [62] for document classification (neoplasm, inflammation).

Despite the annotated data sets, which are essential for applying language technologies, there exist frameworks and tool sets that can be used in this domain. The most essential ones are shortly described in the following. In the past ten years UIMA has become a de facto standard for processing unstructured data sources. It is a general architecture for processing unstructured content like images or sound, but is mostly used for building modular NLP components and aggregating problem domain specific NLP pipelines. UIMA-based processing environments with a special focus on processing clinical narratives are cTakes [40], MedKATp [63] and The Leo framework - The VINCI-developed NLP infrastructure [41, 64]. Originally introduced by IBM [65], it is an open source Apache project with a rich user community. Within the German speaking community, companies and institutions that build on the UIMA framework include Averbis GmbH [66] as a company and the JULIE Lab [67] in Jena Germany as an academic institution. Another framework that has been applied to numerous medical extraction scenarios was MedLEE, developed by Friedman et al. [42, 43]. MetaMap [68] annotates textual data with concepts of the UMLS metathesaurus and is used as core engine for SemRep [69], a system that generates subject-predicate-object triples out of the MetaMap annotations. Usually applied in the biomedical domain its applicability to clinical narratives had been investigated by Liu et al. [70]. HITEx [71] introduced by Zeng et al. [72] is based on GATE [73], a NLP engine. Further NLP engines are openNPL [74], LingPipe [75], and Mallet [76], especially for using Java based development approaches.

Chapter 3

Methods

3.1 Support vector machines

Support Vector Machines (SVMs) [77–79] use the following decision function for classifying instance label pairs $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, l$ for all $\mathbf{x}_i \in \mathbb{R}^n$ to a target value $y \in \{1, -1\}$:

$$
\begin{aligned}
f(\mathbf{x}) &= sgn(\mathbf{w}^T \phi(\mathbf{x}) + b) && (3.1)\\
&= sgn(\sum_{i=1}^{l} \alpha_i y_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) + b) && (3.2)\\
&= sgn(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b) && (3.3)
\end{aligned}
$$

$\mathbf{w} \in \mathbb{R}^n$ being a weight coefficient term and $b \in \mathbb{R}$ defining a bias. For finding the optimal α_i the following minimization problem must be solved [80]:

$$
\begin{aligned}
min_{\mathbf{w},b,\xi} &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i && (3.4)\\
subject\ to &= y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 && (3.5)
\end{aligned}
$$

$\xi_i \in \mathbb{R}$ defines an upper error bound and $C \in \mathbb{R}$ is a tradeoff parameter between the error and margin. Due to the fact that after applying the Lagrangian the final optimization problem depends on the inner product in the form of $\mathbf{x}_i^T\mathbf{x}_j$ one can use the so-called kernel trick, getting the inner product of a kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ without actually performing the real transformation from the input feature space into a higher dimensional one. The major effect is that instances that are not linearly separable in the input space become linearly separable in the higher dimensional feature space, e.g. the XOR problem [78, 81]. SVMs are used as core method in Chapter 4 for clinical document classification and in Chapter 7 in a clinical NLP pipeline.

3.2 Vector space model

The Vector Space Model (VSM) is a general model where observations with afore defined expressions (features) can be represented [82, 83]. This model is especially applied to statistical retrieval or classification approaches using text documents. In this case documents are seen as bag of words and the corresponding terms to a document can be mapped to a unique point in the VSM.

A document collection D can therefore be modeled as consisting of $\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_j, \ldots \mathbf{d}_n$ documents, representing the observations. The entire document collection D forms a vocabulary with $t_1, t_2, t_i, \ldots t_m$ *different* terms or features. The document $\mathbf{d}_j$ can be spatially interpreted as a point in the m-dimensional vector space, respectively forming a m-dimensional feature vector. In this sense a VSM can be described via a $m \times n$ matrix $\mathbf{X}$. In the classic VSM the term-specific weights within the feature vectors are products of local and global parameters. The weights used in this work are described in the next section. VSMs are used in Chapter 4 for clinical document classification and in Chapter 5 in a clinical IR scenario.

3.2.1 Weighting schemes

$$binary : w_{ij} = \begin{cases} 1 & t_i \in \mathbf{d}_j \\ 0 & t_i \notin \mathbf{d}_j \end{cases} \rightarrow \mathbf{d}_j = (0, 1, 1, 0, ..., 1) \tag{3.6}$$

$$tf : w_{ij} = 1 + \log f_{ij} \tag{3.7}$$

$$tf - idf : w_{ij} = \begin{cases} (1 + \log f_{ij}) \cdot \log \frac{N}{n_i} & \text{if } f_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

According to the formal notion of Baeza-Yates et al. [12], f_{ij} counts the occurrences of term t_i within document $\mathbf{d}_j$, N reflects the number of documents within the collection, and n_i the overall frequency of term t_i in the document collection D. In addition to these classical weighting schemes the fourth one was chosen for having the highest significant impact on retrieval performance based on an evaluation made by Abdou and Savoy [84]. This last one I(n)B2 is based on the *Divergence from Randomness* framework [85], and resulted in an 170% enhancement in MAP compared to the tf-idf weighting scheme when applied to retrieval tasks using annotated MEDLINE documents from the TREC 2004 text retrieval conference [86]. The following description is based on Abdou and Savoy [84]. I(n)B2 combines the information measures:

$$w_{ij} = Inf_{ij}^1 \cdot Inf_{ij}^2 = -log_2[Prob_{ij}^1] \cdot (1 - Prob_{ij}^2) \tag{3.9}$$

$Prob_{ij}^1$ expresses the likelihood to find tf_{ij} counts within the document $\mathbf{d}_j$ of the term t_i by pure chance. The probability to find another occurrence in the document $\mathbf{d}_j$ of the term t_i, with the condition that tf_{ij} occurrences were already found is denoted via $Prob_{ij}^2$.

I(n)B2 can be expressed by following formula now:

$$Prob_{ij}^{1} = \left(\frac{df_i + 0.5}{n+1}\right)^{tfn_{ij}} \tag{3.10}$$

$$Prob_{ij}^{2} = 1 - \left(\frac{tc_i + 1}{df_i \cdot (tfn_{ij} + 1)}\right) \tag{3.11}$$

$$tfn_{ij} = tf_{ij} \cdot log_2\left(1 + \frac{c \cdot \text{mean } dl}{l_j}\right) \tag{3.12}$$

The amount of documents indexed with the term t_i are described by df_i. tc_i indicates the overall occurrence of term t_i within the document collection, n accounts for the absolute number of corpus documents with mean dl representing the average document length and finally a constant $c = 1.5$ is applied. The four weighting schemes are used in Chapter 4 for clinical document classification, and in Chapter 5 in a clinical IR scenario.

3.2.2 Similarity measure

Within the VSM different similarity measures between documents $\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_j, \ldots \mathbf{d}_n$ or a mapped query $\mathbf{q}$ to its surrounding documents can be applied. One similarity measure often used in this space is the so-called cosine similarity, which is calculated between a document $\mathbf{d}_j$ and a query $\mathbf{q}$ as follows:

$$sim(\mathbf{d}_j, \mathbf{q}) = \frac{\mathbf{d}_j \cdot \mathbf{q}}{|\mathbf{d}_j|\,|\mathbf{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{3.13}$$

The similarity measure is used in Chapter 5 in a clinical IR scenario.

3.3 Latent semantic analysis

Latent Semantic Analysis (LSA) [87, 88] is a statistical retrieval method exploiting term co-occurrences within a term-document matrix and is generally categorized as a kind of *distributional semantics*.

m terms and n documents form a sparse $m \times n$ matrix $\mathbf{X}$, the afore explained VSM in Section 3.2. Different weighting schemes can be applied at this stage to the term-document matrix as described in the Section 3.2.2. The core of LSA is to apply a Singular Value Decomposition (SVD) on the term-document matrix $\mathbf{X} = \mathbf{TSD}^T$ getting the orthonormal matrices $\mathbf{T}$ and $\mathbf{D}^T$ with the eigenvectors of $\mathbf{XX}^T$ and $\mathbf{X}^T\mathbf{X}$. $\mathbf{T}$ is often called the term matrix and $\mathbf{D}^T$ the document matrix. The roots of the eigenvalues of $\mathbf{XX}^T$ and $\mathbf{X}^T\mathbf{X}$ are embedded in $\mathbf{S}$.

The degree of dimensionality reduction can be controlled by eliminating the lowest eigenvalues and their eigenvectors to a new dimension k (Figure 3.1). A query $\mathbf{q}$, also called a pseudo document, can be transformed into the document space via:

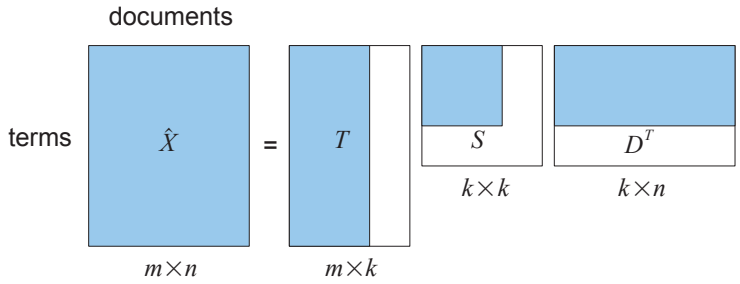


Figure 3.1: Singular value decomposition and dimensionality reduction of the the original term document matrix [89].

$$\mathbf{Q}_k = \mathbf{q}^T \mathbf{T}_k \mathbf{S}^{-1} \tag{3.14}$$

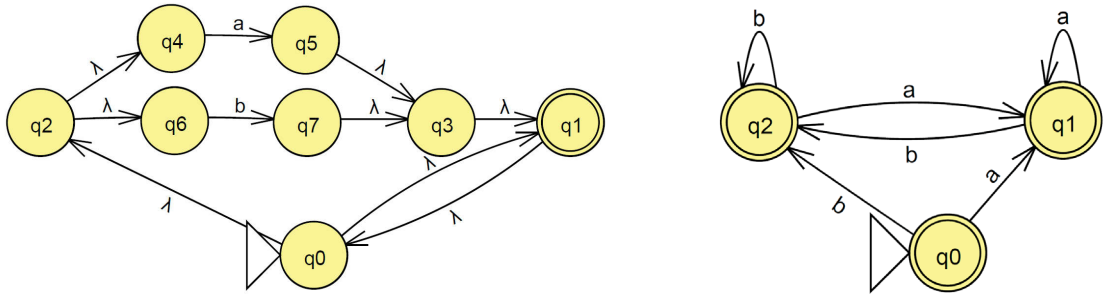
$\mathbf{Q}_k$ is in the same semantic space as $\mathbf{D}^T$ now. By use of the cosine similarity (Section 3.2.2), documents can be ranked to the initial query $\mathbf{Q}_k$ in this semantic space with the degree of dimensionality reduction k. LSA is used in Chapter 5 in a clinical IR scenario.

3.4 Regular expressions

Since its formalization in the late 1950s by the mathematician Stephen Kleene [90], the concept of regular expressions has been a highly useful asset in computer engineering. Regular expressions are typically applied in IR or IE and belong to the class of *rule-based* systems, in contrast to systems based on *machine learning*. With a regular expression one can express a subset of allowed character sequences within a set of all possible ones, defining a so called pattern. Regular expression patterns include (i) meta-characters each of which has special functionality or meaning within the regular expression and (ii) regular characters providing the literal meaning. In case meta-characters are used with their literal meaning, they are marked according to the regular expression syntax in use. The formulated regular expression can be processed to define a NFA [91, 92] and its isomorph a DFA through the subset construction algorithm [93, 94]. An example of both automaton types is depicted in Figure 3.2. To complete this section, the 5-tuple definition $(Q, \sum, \delta, q_o, F)$ of a NFA is given according to Sipser [94] with:

1. Q is a finite set of states,
2. $\sum$ is a finite alphabet,
3. $\delta : Q \times \sum_\lambda \longrightarrow P(Q)$ is the transition function,
   with $P(Q)$ being then power set of Q and $\sum_\lambda = \sum \cup \{\lambda\}$
4. $q_0 \in Q$ is the start state, and
5. $F \subseteq Q$ is the set of accepted states

The description of the DFA differs from the one of the NFA just in the type of the transition function. The transition function of the NFA takes a state, input symbol or the empty string to define the set of possible next states. In contrast to the NFA the DFA takes a state, an input symbol and produces the next state. A more detailed description as well as the formal proof that out of any NFA a DFA can be constructed is given in [94]. In this work a Java based regular expression engine was used. A detailed description of the syntax which can be used to form the regular expression can be found in the official Java 7 API [95]. Regular expressions are exploited in Chapter 6 in a clinical IE scenario.



(a) Nondeterministic finite automaton representation

(b) Deterministic finite automaton representation

Figure 3.2: NFA and DFA representation of the regular expression `(a|b)*` using JFLAP [96]

3.5 Morphosemantic processing

The MorphoSaurus system [97] is a rule based system where subwords $\mathcal{S}$ (mostly semantically atomic word fragments such as morphemes) are formally represented by a quadruple $S \subset \mathcal{SID} \times \mathcal{T} \times \mathcal{L} \times \mathcal{D}$. $\mathcal{SID}$ defines a set of allowed subwords. $\mathcal{T}$ represents the type of the subword out of seven defined ones: subword stems (ST) e.g. “gastr”, “hepat”, “head”; prefixes (PF) e.g. “de-”, “hyper-”, “anti-”; proper prefixes (PP) e.g. “peri-”, “hemi-, “down-”; infixes (IF) e.g. “-o-” or “-r-”; suffixes (SF) e.g. “-a”, “-tomy”, “-itis”; proper suffixes (PS) e.g. “-ing”, “-ieron” and invariantes (IV) e.g. “ion”. $\mathcal{L}$ defines a set of allowed languages and $\mathcal{D}$ specifies the domain of the subword.

The application of the so-called morphosemantic indexing consists of three steps depicted in Figure 3.4:

Orthographic normalization. The whole text is transformed into lower letters as well as certain characters are replaced and normalized e.g. “ä, ö, ü, ß” to “ae, oe, ue, ss”, according to language-specific transformation rules.

Morphological segmentation. Through the use of a subword lexicon that is specific to a language and a domain, the text is transformed into a sequence of subwords. Allowed subword decompositions are determined via the automaton depicted in Figure 3.3.

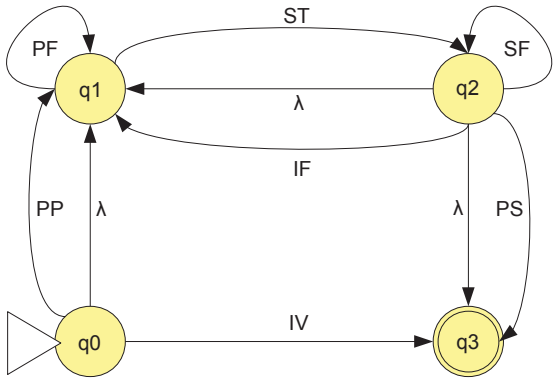


Figure 3.3: NFA representation of allowed subword type sequences [98].

Semantic normalization. In this step subwords are mapped to the language independent final morphosaurus identifiers, which represent a kind of domain concepts. "Myokarditis": #muscle #heart #inflamm; "Herzmuskelentzündung": #heart #muscle #inflamm; "Inflammation of the heart muscle": #inflamm #heart #muscle [97];

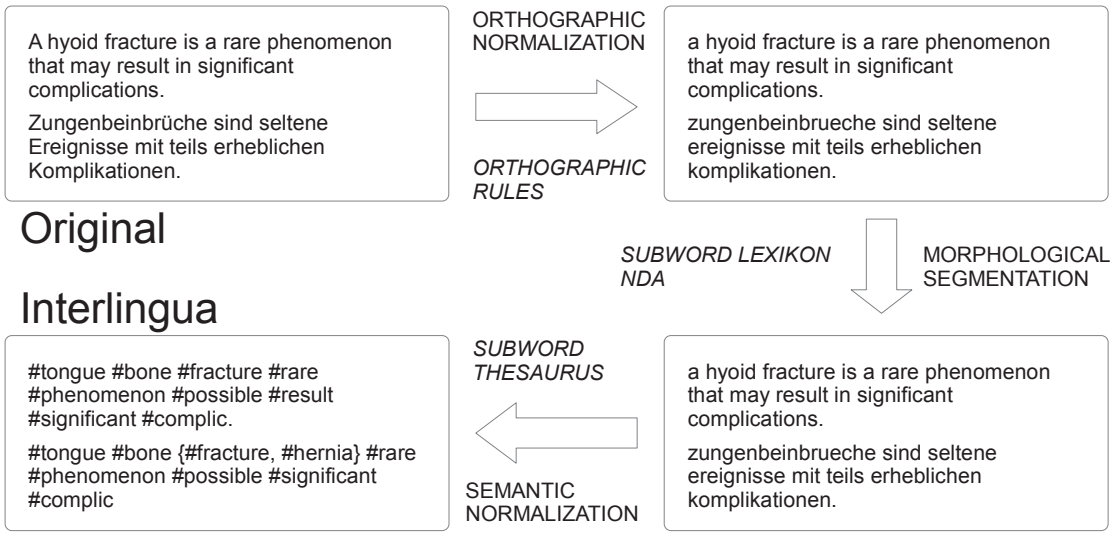


Figure 3.4: Processing scheme of morphosemantic indexing [98].

Morphosemantic indexing is used in Chapter 4 for clinical document classification, and in Chapter 5 in a clinical IR scenario.

3.6 Evaluation metrics

This section describes in detail the evaluation metrics used to estimate the performance of the different language technology approaches. The mathematical notation is based on the description of Manning et al. [11].

Precision (P, positive predictive value) is the probability that a retrieved item is relevant.

$$Precision = \frac{\#\text{ (relevant items retrieved)}}{\#\text{ (retrieved items)}} = P(relevant|retrieved) \tag{3.15}$$

Recall (R, hit rate, sensitivity, true positive rate) is the probability that a relevant item was found.

$$Recall = \frac{\#\text{ (relevant items retrieved)}}{\#\text{(relevant items)}} = P(retrieved|relevant) \tag{3.16}$$

F-measure is the weighted harmonic mean of recall and precision.

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R})} = \frac{(\beta^2+1)PR}{\beta^2 P + R},\ \beta^2 = \frac{1-\alpha}{\alpha} \tag{3.17}$$

Setting $\beta = 1$ results in:

$$F_{\beta=1} = \frac{2PR}{P+R} = F_1 \tag{3.18}$$

These are the most prominent evaluation measures widely used for evaluating IR and IE systems. In the latter case there is also the distinction between exact and inexact matches regarding annotation borders. If the IE system annotates a text within the same window like in the gold standard then an exact match is given. If there is a partial match it is referred to as an inexact match. Further, there is a distinction between micro and macro-averaging of extraction or classification results. For micro-averaging the confusion tables ($n \times n$ table, predicted versus actual class) are added together; and precision, recall and F-measure values refer to this table. Macro-averaging calculates, per confusion matrix, precision, recall and F-measure, adds the results for all extraction or classification results and divides the added up evaluation measures by the number of extraction/classification tasks. While micro-averaging has the advantage to get more influence from the most common classes, macro averaging gives each class the same priority without considering the amount of elements it contains. Despite the fact that the aforementioned evaluation results belong to the class of unranked metrics, for ranked retrieval results within the domain of IR evaluation in this work Precision at k (P_k corresponds to the precision when k documents are retrieved) and MAP is used, which is defined for a set Q of information needs as follows:

$$MAP(Q) = \frac{1}{Q}\sum_{j=1}^{|Q|} AP(q_j) \tag{3.19}$$

The formula expresses the mean of the Average Precision (AP) of information needs $q_j \in Q$. The AP for one information need q_j is defined as:

$$AP(q_j) = \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}) \tag{3.20}$$

R_{jk} is the subset of ranked retrieval results until one gets to the document d_k within the set of relevant documents $d_1, ...d_{m_j}$ for an information need $q_j \in Q$. AP can be interpreted as the non-interpolated area under the precision/recall curve (per retrieved document, recall and precision are plotted on a 2-D plot forming saw-tooth shape). MAP can be seen as the averaged area of all precision/recall graphs for a set of information needs Q.

For the evaluation and training of classifiers using machine-learning methods, 10-fold-cross validation is widely applied [7]. The data set under investigation is split into ten equal sets. Nine sets together form the training set on which the classifier is trained, and the remaining set is used for validation. This process is repeated 10 times, with each of the 10 sets used exactly once as the validation data. Typically, out of the 10 resulting confusion tables a micro-averaged F-measure is calculated, which estimates the performance of the classifier under test. In this work, a chi-squared test was applied to test significant performance differences. The higher the number of items used for this test, the more likely it signals a significant difference. A relevant difference for machine learning methods was defined when $\Delta F_1 \geq 0.01$, micro-avg. $\Delta F_1 \geq 0.01$ or macro-avg. $\Delta F_1 \geq 0.01$.

Chapter 4

Clinical Document Classification

4.1 Introduction

The dense information within clinical texts leads to a problem commonly known as information overload [99]. E.g., a patient record of a 2 1/2 year old child may contain more than 300 documents [100]. No clinician has the time to read all this in detail. Optimizing the scarce time for getting the information that is really relevant is a non-trivial task. The problem can be alleviated by pre-classifying the documents. E.g., it has proved very helpful to classify pathology findings reports according to the mention of the existence of an inflammation or neoplasm. This supports the process of fast diagnostic reporting (navigational and from a visual perception point of view). A rule-based system has been implemented by Schmiedberger and Errath [62] and is an integral part of the CIS [61, 101].

This chapter is about the use of machine learning methods in combination with morphosemantic processing for document classification. It is organized as follows: Section 4.2 provides a literature survey in the context of classifying clinical narratives and motivates our investigation and chosen methods. Section 4.3 describes the methods, data sets and evaluation metrics in detail. Section 4.4 presents and discusses the evaluation results. The final Section 4.5 discusses possible follow-up investigations.

4.2 Related work

This literature review is about classification methods applied to clinical narratives. It excludes approaches that concentrate on biomedical publications or other proof-read textual data domains. It focuses on cases where complete documents are classified into one or more categories. This distinction is crucial, because machine learning methods can also be used to classify text passages within a document, e.g. for NER or concept mapping, in contrast to document classification methods where a whole document gets annotated.

Wilcox and Hripcsak [102] tested 6 different machine learning algorithms (naive Bayes, decision tables, instance-based inducer, MC4, C5.0, CN2) and their performance on 200 chest X-ray reports. The MedLEE NLP engine was used for text preprocessing and feature

generation. The presence of six different clinical conditions were tested: congestive heart failure, chronic obstructive pulmonary disease, acute bacterial pneumonia, neoplasm, pleural effusion without congestive heart failure, and pneumothorax. CN2 preformed best (expert feature selection: specificity 0.99, sensitivity 0.48). Decision tables had a significantly lower performance compared to the other classifiers except naive Bayes. Dimensionality reduction was performed via bootstrapping, but top-down feature selection done by domain experts led to a better classification performance with regard to sensitivity. A higher number of positive cases in the training set was associated with improved classification performance. The authors concluded that the performance of machine learning is not as high as rules written by a domain expert.

McCowan et al. [103, 104] classified lung cancer stages according to the Tumor Node Metastases (TNM) classification developed by the American Joint Committee on Cancer (AJCC) and International Union against Cancer (IUC). The corpus was built from two sources including 718 lung cancer patients: narratives containing pathologic staging decisions for lung cancer patients (Queensland Integrated Lung Cancer Outcomes Project (QILCOP) data) and histology reports for lung cancer patients from the state pathology information system (AUSLAB). Text preprocessing included normalization of acronyms, numbers and dimensions, expansion of abbreviations, and identification of spelling variants, annotation with UMLS concept identifiers and detection of negated expressions. A VSM was build using the log-tf-idf-cosine (LTC) weighting scheme. For statistical classification a SVM (SVM^{light}) was trained and evaluated (100 folds cross validation) for each TNM stage category (T1-T4, N0-N2). McCowan et al. [103] reported a micro-average F-measure of 0.61 for T staging and a micro-average F-measure of 0.78 for N staging, respectively. The Receiver Operator Characteristic (ROC) curve showed an Area Under the Curve (AUC) of 0.86 for T and 0.82 for N staging. The approach was further expanded for multi-class classification using SVMs by Nguyen et al. [105]. Four different architecture types were therefore evaluated: maximum normalized score, hierarchy of binary SVM's, multi-class SVM and multi-class SVM on binary outputs. The hierarchical approach showed the most promising results achieving an overall accuracy of 0.64 and 0.82 across T and N stages.

Jouhet et al. [106] compared two different classifiers for automatically categorizing pathology reports. A gold standard was set up comprising 5,121 manually annotated pathology reports. The classifiers under scrutiny were a Naive Bayes one and a SVM in combination with different weighting schemes (term frequency, term frequency-inverse document frequency, term frequency-inverse class frequency) to setup a VSM. Preprocessing contained stemming and stop-word removal. The best performance was achieved using the SVM in combination with the term-frequency inverse class frequency weighting scheme. International Agency for Research on Cancer (IARC) categories were applied with an F-measure of 0.97 for both topography and morphology. ICD-0-3 codes were assigned with an F-measure of 0.72 for topography and 0.85 for morphology.

Hiissa et al. [107, 107] classified 1,363 text snippets of intensive care nursing narratives into three distinct classes: Breathing, Blood Circulation and Pain. For this purpose, the narratives were annotated by three medical domain experts with a good inter-rater reliability (Cohen's kappa greater than 0.8). The text fragments were manually split into a training (n=708) and test (n=655) set. A Least Squares SVM was trained and further evaluated on the test set. The performance was evaluated via AUC. The average

performance of the classification approach for all three classes was about 0.85. This is quite remarkable as the machine learning approach reached similar values as the human manual annotators performing the same task.

Spat et al. [108] applied a multi-label text classification system to German clinical narratives. To this end, 1,500 documents were annotated using the following categories, allowing also multiple labels per document: radiology and physiotherapy, neurology, anesthesia and intensive care, internal medicine, vascular surgery, casualty surgery, and surgery. Two scenarios were compared, *viz.* with text preprocessing and without. The results showed that J48 performed best with an additional preprocessing of the texts, reaching an F-measure of 0.89. Without preprocessing the evaluation measures were slightly lower. In a second step, the trained classifier was used in a prototypical medical IR system (MIRS) to boost the ranking of retrieved documents according to a chosen category [109].

Shiner et al. [110] used two different machine learning algorithms (Conditional Random Fields (CRFs) and Maximum Entropy (MaxEnt)) to determine whether a psychiatric note from a gold standard of 221 notes, can be classified as "individual psychology" according to the guidelines of Dieperink et al. [111]. They showed on the one hand that administrative data overestimate the number of psychiatric sessions of U.S. veterans, and on the other hand that the MaxEnt classifier performed well in automatically classifying the notes via the Automated Retrieval Console (ARC) tool, reaching an F-measure of 0.93 (recall 0.97, precision 0.93).

Afzal et al. [112] applied four machine learning methods on epidemiological case identification: the decision-tree learners C4.5 and MyC, a SVM, and the decision-rule learner RIPPER. Two annotated sets were built for this purpose, one about hepatobiliary disease (656 positive and 317 negatively labeled patients), and a second with acute renal failure (237 positive and 3,751 negatively labeled patients). The aforementioned classifiers were trained and tested using a five-fold cross-validation varying the degree of over-sampling, under-sampling, cost sensitive learning, and the imbalance ratio (positive to negative cases). A chi-squared feature selection was performed. Performance values were in-between a maximum sensitivity of 0.95 and minimum specificity of 0.54 for both classification tasks. C4.5 and MyC showed the best overall performance.

Patterson et al. [41] compared a rule-based document classification approach against three different machine learning methods to identify colonoscopy reports documented for screening purposes. A gold standard with 3000 (training set 2000, test set 1000) annotated colonoscopy reports with the following classes was created: "screening", "non-screening", "non-colonoscopy" and "unknown". SVM-based classification without feature selection performed best on the training set with an accuracy of 0.995. Test set based evaluation showed that the rule based knowledge engineering method outperformed the machine learning approaches with an accuracy of 0.95 compared to the best machine learning model, a SVM with feature selection, reaching an accuracy of 0.94.

The study in this chapter investigates to what extent machine learning methods can be applied to the task of classifying pathology narratives into four distinct classes. For this purpose, an existing gold standard is used to explore how well parameter-optimized machine learning methods in combination with different preprocessing methods and text representation forms can be applied. The hypothesis is that performance levels with an F-

measure greater than 0.95 can be reached for the defined classification task in the proposed evaluation setting with machine learning methods. In addition, the impact of different preprocessing methods on the classification performance is investigated and discussed in the light of a possible implementation within in a clinical environment, considering real time constraints and classification performance. To the best of the author's knowledge, this is the first approach of this kind applied to German clinical narratives. The setting and methods are described in detail in the next section.

4.3 Materials and methods

4.3.1 Overview

This study used a SVM exploiting a linear kernel (Section 3.1) due to its known good performance for classifying textual content [5]. The term-document matrix built for feature space generation was based on three different preprocessing methods. One time, the narratives were preprocessed with the Lucene SnowballAnalyzer applying a German stop word list (Appendix A.1). The second time, morphosemantic analysis (Section 3.5) was applied to the text base under scrutiny. The third textual representation was a merged variant from both preprocessing outputs (Section 4.3.3). In a next step, different character n-gram (substrings of length 2-5) variations were generated out of the representations forms explained above. The different textual representation variants of the original narratives were used for feature space generation (term-document matrix generation) then. Four different weighting schemes were applied on the term-document matrix: binary (bin), term frequency (tf), term frequency-inverse document frequency (tf-idf) and I(n)B2 as described in Section 3.2.1. The different generated feature spaces were finally used for training and applying a 10 fold cross validation per classification task (i) *Inflammation* yes/no , (ii) *Neoplasm* yes/no. The gold standard needed for this evaluation is described in the next section.

4.3.2 Gold standard

To have a representative sample size that reflects the diversity of the document under investigation ∼3,000 pathology were manually annotated by two domain experts. This sample size also corresponds to one of the most recent publication in the area of clinical document classification by Patterson et al. [41]. If annotations between two domain experts were different, a third expert was included to reach a consensus. Two main annotation categories were applied for neoplasm and inflammation for each pathology report [61, 62]. For neoplasm the fields "Organ" (topography code), "Group" (morphology code group range), "Morphology" (morphology code) and "Dignity" (morphology code, behavior digit) were annotated out of the International Classification of Diseases for Oncology - Revision Three (ICD-O-3) system. Additionally, another two annotation fields were added with a defined set of possible values. The first one added was whether a neoplasm was mentioned at all. The second field contained a confidence value about the annotations itself, using the values set: 1 $\widehat{=}$ "probably", 2 $\widehat{=}$ "compatible with", 3 $\widehat{=}$ "questionable", 4 $\widehat{=}$ "possible",

5 $\widehat{=}$ "uncertain", 6 $\widehat{=}$ "unlikely", 7 $\widehat{=}$ "safe", 8 $\widehat{=}$ "certainly not" and case 9 $\widehat{=}$ "not mentioned". For inflammation a topography code was assigned, followed by a confidence value out of the above value set, as well as whether an inflammation was mentioned at all.

4.3.3 Data

The following examples depict the output of the preprocessing steps mentioned above (Section 4.3.1) to highlight the differences in the textual information representation forms. Besides the typical idiosyncrasies of clinical narratives (Section 2.2), these ones suffer from an additional circumstance, *viz.* that all documents are in upper case due to encoding restrictions of a legacy documentation system, see below:

Original (*Orig*):

> OBERFLAECHLICHE ANTEILE EINES VILLOESEN DICKDARMSCHLEIMHAUTADENOMS (MITTELGRADIGE DYSPLASIE): WHO: LOW GRADE INTRAEPITHELIALE NEOPLASIE. DAS VERHALTEN ZUR UNTERLAGE AUS DER OBERFLAECHLICHEN BIOPSIE NICHT BEURTEILBAR. IM VORLIEGENDEN MATERIAL KEINE SCHWERE DYSPLASIE.

Lucene SnowballAnalyzer (*Snowball*):

> oberflaech anteil villoes dickdarmschleimhautadenom mittelgrad dysplasi who low grad intraepithelial neoplasi verhalt unterlag oberflaech biopsi beurteilbar vorlieg material schwer dysplasi

Morphosemantic normalization (*Morpho*):

> superficiiiwiky componentiiixjx villiiiqqkr coloniciiiwij mucosaiikrjz adenomaipypqw medialiikpww instrumentiiiwkk medicamentiiikpx graduateiijjyx dysplasipxrrr healthiijjiw organizijipxx worldijyyrq fewikrizi graduateiijjyx internaliiiqij epitheliiixyr neoplastiikxpp behaviouriiwpwy fundamentipwzkx documentationijirzz superficiiiwiky extractiiiywr histioijixij notrjryrz estimatiiiypx beingiiziyp substanciikpqy withoutipqkqx heavyrijyik dysplasipxrrr

SnowballAnalyzer ∪ Morphosemantic normalization (*Mixed*):

> oberflaech anteil villoes dickdarmschleimhautadenom mittelgrad dysplasi who low grad intraepithelial neoplasi verhalt unterlag oberflaech biopsi beurteilbar vorlieg material schwer dysplasi superficiiiwiky componentiiixjx villiiiqqkr coloniciiiwij mucosaiikrjz adenomaipypqw medialiikpww instrumentiiiwkk medicamentiiikpx graduateiijjyx dysplasipxrrr healthiijjiw organizijipxx worldijyyrq fewikrizi graduateiijjyx internaliiiqij epitheliiixyr neoplastiikxpp behaviouriiwpwy fundamentipwzkx documentationijirzz superficiiiwiky extractiiiywr histioijixij notrjryrz estimatiiiypx beingiiziyp substanciikpqy withoutipqkqx heavyrijyik dysplasipxrrr

3-gram example of the SnowballAnalyzer ∪ Morphosemantic normalization:

> obe ber erf rfl fla lae aec ech ant nte tei eil vil ill llo loe oes dic ick ckd kda dar arm rms msc sch chl hle lei eim imh mha hau aut uta tad ade den eno nom mit itt tte tel elg lgr gra rad dys ysp spl pla las asi who low gra rad int ntr tra rae aep epi pit ith the hel eli lia ial neo eop opl pla las asi ver erh rha hal alt unt nte ter erl rla lag obe ber erf rfl fla lae aec ech bio iop ops psi beu eur urt rte tei eil ilb lba bar vor orl rli lie ieg mat ate ter eri ria ial sch chw hwe wer dys ysp spl pla las asi sup upe per erf rfi fic ici cii iii iiw iwi wik iky com omp mpo pon one nen ent nti tii iii iix ixj xjx vil ill lli lii iii iiq iqq qqk qkr col olo lon oni nic ici cii iii iiw iwi wij muc uco cos osa sai aii iik ikr krj rjz ade den eno nom oma mai aip ipy pyp ypq pqw med edi dia ial ali lii iik ikp kpw pww ins nst str tru rum ume men ent nti tii iii iiw iwk wkk med edi dic ica cam ame men ent nti tii iii iik ikp kpx gra rad adu dua uat ate tei eii iij ijj jjy jyx dys ysp spl pla las asi sip ipx pxr xrr rrr hea eal alt lth thi hii iij ijj jji jiw org rga gan ani niz izi zij iji jip ipx pxx wor orl rld ldi dij ijy jyy yyr yrq few ewi wik ikr kri riz izi gra rad adu dua uat ate tei eii iij ijj jjy jyx int nte ter ern rna nal ali lii iii iiq iqi qij epi pit ith the hel eli lii iii iix ixy xyr neo eop opl pla las ast sti tii iik ikx kxp xpp beh eha hav avi vio iou our uri rii iiw iwp wpw pwy fun und nda dam ame men ent nti tip ipw pwz wzk zkx doc ocu cum ume men ent nta tat ati tio ion oni nij iji jir irz rzz sup upe per erf rfi fic ici cii iii iiw iwi wik iky ext xtr tra rac act cti tii iii iiy iyw ywr his ist sti tio ioi oij iji jix ixi xij not otr trj rjr jry ryr yrz est sti tim ima mat ati tii iii iiy iyp ypx bei ein ing ngi gii iiz izi ziy iyp sub ubs bst sta tan anc nci cii iik ikp kpq pqy wit ith tho hou out uti tip ipq pqk qkq kqx hea eav avy vyr yri rij ijy dys ysp spl pla las asi sip ipx pxr xrr rrr

4.4 Results and discussion

4.4.1 Neoplasm detection

The results for neoplasm classification are depicted in Table 4.1. The maximum achieved performance value using a 10-fold cross-validation has a micro-averaged F-measure of 0.96 using a mixed text-base with 3-gram decomposition (Section 4.3.3) in combination with the I(n)B2 weighting scheme (Section 3.2.1). The results based on this table are discussed in more detail in the following sections starting with the investigation what impact a varying character n-gram representation has on the classification results.

As depicted in Table 4.1 using the SnowballAnalyzer alone, 4-gram and 3-gram decomposing have a positive significant impact on the evaluation performance on most of the weighting schemes. Interestingly, when using the mixed representation form none of the character n-gram schemes have an impact, except the 2-gram scheme was found. In this case, a negative impact can be shown for all weighting schemes. The same observation was made using morphosemantic normalization alone.

	Snowball				Mixed				Morpho			
	bin	tf	tf-idf	I(n)B2	bin	tf	tf-idf	I(n)B2	bin	tf	tf-idf	I(n)B2
term	0.937'	0.936'	0.932'	0.938'	0.957'	0.959'	0.954'	0.955'	0.950'	0.951'	0.945'	0.950'
5-gram	0.947	0.947	0.944	0.947	0.956	0.959	0.951	0.959	0.949	0.948	0.943	0.952
4-gram	0.950*	0.950*	0.947*	0.954*	0.956	0.959	0.953	0.959	0.947	0.949	0.943	0.951
3-gram	0.948	0.951*	0.947*	0.956*	0.959	0.957	0.952	0.962	0.948	0.950	0.943	0.954
2-gram	0.928	0.932	0.930	0.935	0.923*	0.939*	0.937*	0.937*	0.917*	0.931*	0.928*	0.924*

Table 4.1: Micro-averaged F-measure 10-fold cross-validation results for the category *Neoplasm*. Evaluation results marked with * exhibit a significant difference to the baseline (term) marked with ' ($p < 0.05$). The maximum value is underpinned in gray.

Surprisingly, most of the time varying weighting schemes had no impact on the classification task with fixed character n-gram decomposition and preprocessing scheme, see Table 4.2. Using 2-grams in combination with the category *Mixed*, more advanced weighting schemes (tf, tf-idf and I(n)B2) had a positive impact compared to a simple binary weighting approach. For morphosemantic normalization alone just the weighting scheme tf had a significant positive impact. In all other cases the weighting scheme had no significant performance influence.

	term	5-gram	4-gram	3-gram	2-gram
Snowball					
bin	0.937'	0.947'	0.950'	0.948'	0.928'
tf	0.936	0.947	0.950	0.951	0.932
tf-idf	0.932	0.944	0.947	0.947	0.930
I(n)B2	0.938	0.947	0.954	0.956	0.935
Mixed					
bin	0.957'	0.956'	0.956'	0.959'	0.923'
tf	0.959	0.959	0.959	0.957	0.939*
tf-idf	0.954	0.951	0.953	0.952	0.937*
I(n)B2	0.955	0.959	0.959	0.962	0.937*
Morpho					
bin	0.950'	0.949'	0.947'	0.948'	0.917'
tf	0.951	0.948	0.949	0.950	0.931*
tf-idf	0.945	0.943	0.943	0.943	0.928
I(n)B2	0.950	0.952	0.951	0.954	0.924

Table 4.2: Micro-averaged F-measure 10-fold cross-validation results for the category *Neoplasm*. Evaluation results marked with * exhibit a significant difference to the baseline (bin) marked with ' ($p < 0.05$). The maximum value is underpinned in gray.

Table 4.3 shows that morphosemantic processing without character n-gram decomposing, always had a positive noticeable performance impact regardless of the weighting scheme in use. Twice, using the tf and I(n)B2 weighting schemes, this performance gain also occurred when using 5-grams.

	term	5-gram	4-gram	3-gram	2-gram
bin					
Snowball	0.937’	0.947’	0.950’	0.948’	0.928’
Mixed	0.957*	0.956	0.956	0.959	0.923
Morpho	0.950*	0.949	0.947	0.948	0.917
tf					
Snowball	0.936’	0.947’	0.950’	0.951’	0.932’
Mixed	0.959*	0.959*	0.959	0.957	0.939
Morpho	0.951*	0.948	0.949	0.950	0.931
tf-idf					
Snowball	0.932’	0.944’	0.947’	0.947’	0.930’
Mixed	0.954*	0.951	0.953	0.952	0.937
Morpho	0.945*	0.943	0.943	0.943	0.928
I(n)B2					
Snowball	0.938’	0.947’	0.954’	0.956’	0.935’
Mixed	0.955*	0.959*	0.959	0.962	0.937
Morpho	0.950*	0.952	0.951	0.954	0.924

Table 4.3: Micro-averaged F-measure 10-fold cross-validation results for the category *Neoplasm*. Evaluation results marked with * exhibit a significant difference to the baseline (Snowball) marked with ’ ($p < 0.05$). The maximum value is underpinned in gray.

4.4.2 Inflammation detection

	Snowball				Mixed				Morpho			
	bin	tf	tf-idf	I(n)B2	bin	tf	tf-idf	I(n)B2	bin	tf	tf-idf	I(n)B2
term	0.959’	0.953’	0.950’	0.947’	0.973’	0.971’	0.970’	0.970’	0.968’	0.963’	0.962’	0.963’
5-gram	0.965	0.961	0.958	0.964*	0.971	0.971	0.972	0.970	0.967	0.963	0.961	0.965
4-gram	0.970*	0.967*	0.966*	0.967*	0.972	0.970	0.969	0.971	0.968	0.962	0.958	0.966
3-gram	0.967	0.966*	0.966*	0.968*	0.971	0.967	0.963	0.972	0.968	0.962	0.956	0.968
2-gram	0.937*	0.942	0.942	0.945	0.956*	0.951*	0.949*	0.956*	0.949*	0.942*	0.936*	0.952*

Table 4.4: Micro-averaged F-measure 10-fold cross-validation results for the category *Inflammation*. Evaluation results marked with * exhibit a significant difference to the baseline (term) marked with ’ ($p < 0.05$). The maximum value is underpinned in gray.

The results for the classification task *Inflammation* are listed in Table 4.4. It shows a maximum micro-averaged F-measure of 0.97 using a mixed text representation form, without character n-gram decomposing and a binary weighting scheme of the feature vectors. Compared to the maximum performance result of the neoplasm classification task (F-measure 0.96), both results have in common that they rely on a joint representation (category *Mixed*) of the texts under scrutiny. Interestingly, in the case of inflammation classification a simple binary feature weighting scheme without character n-gram decomposing achieved the best result. For the classification of neoplasm the I(n)B2 weighting scheme together with 3-grams produced the top classification performance, however at the cost of much higher computational efforts. In accordance with Table 4.1 character 4-gram decomposi-

tion always had a positive performance impact regardless of the weighting scheme in use, if no morphosemantic component was attached to the processing pipeline. 2-grams models together with a morphological component always had a significant negative performance impact on the classification performance. This is in agreement with the results of neoplasm classification.

What can be seen from Table 4.5 is that in about 95% of the cases the weighting scheme had no significant impact on *Inflammation* classification performance. This corresponds with the results of neoplasm classification listed in Table 4.2.

	term	5-gram	4-gram	3-gram	2-gram
Snowball					
bin	0.959’	0.965’	0.970’	0.967’	0.937’
tf	0.953	0.961	0.967	0.966	0.942
tf-idf	0.950	0.958	0.966	0.966	0.942
I(n)B2	0.947*	0.964	0.967	0.968	0.945
Mixed					
bin	0.973’	0.971’	0.972’	0.971’	0.956’
tf	0.971	0.971	0.970	0.967	0.951
tf-idf	0.970	0.972	0.969	0.963	0.949
I(n)B2	0.970	0.970	0.971	0.972	0.956
Morpho					
bin	0.968’	0.967’	0.968’	0.968’	0.949’
tf	0.963	0.963	0.962	0.962	0.942
tf-idf	0.962	0.961	0.958	0.956*	0.936*
I(n)B2	0.963	0.965	0.966	0.968	0.952

Table 4.5: Micro-averaged F-measure 10-fold cross-validation results for the category *Inflammation*. Evaluation results marked with * exhibit a significant difference to the baseline (bin) marked with ’ ($p < 0.05$). The maximum value is underpinned in gray.

Table 4.6 shows the influence of document preprocessing on the classification result. Without character n-gram variation, morphosemantic normalization merged with an applied Lucene SnowballAnalyzer (*Mixed*) always showed a significant impact on the classification performance. In two cases (tf-idf and I(n)B2), the morphosemantic normalization alone produces a noticeable positive impact. In the remaining cases the categories *Mixed* and *Morpho* had in about 85% no impact on the evaluation performance using character n-gram variation.

4.4.3 Discussion

Both classifications show that the mixed clinical narratives - merging morphosemantic normalization and the Lucene SnowballAnalyzer, always have a positive impact on the classification *if no* character n-gram decomposing is applied. Character n-gram variation together with a morphological component added, has no positive impact on the quality of the classification task. The use of 2-grams together with the morphosemantic nor-

	term	5-gram	4-gram	3-gram	2-gram
bin					
Snowball	0.959’	0.965’	0.970’	0.967’	0.937’
Mixed	0.973*	0.971	0.972	0.971	0.956*
Morpho	0.968	0.967	0.968	0.968	0.949*
tf					
Snowball	0.953’	0.961’	0.967’	0.966’	0.942’
Mixed	0.971*	0.971*	0.970	0.967	0.951
Morpho	0.963	0.963	0.962	0.962	0.942
tf-idf					
Snowball	0.950’	0.958’	0.966’	0.966’	0.942’
Mixed	0.970*	0.972*	0.969	0.963	0.949
Morpho	0.962*	0.961	0.958	0.956*	0.936
I(n)B2					
Snowball	0.947’	0.964’	0.967’	0.968’	0.945’
Mixed	0.970*	0.970	0.971	0.972	0.956
Morpho	0.963*	0.965	0.966	0.968	0.952

Table 4.6: Micro-averaged F-measure 10-fold cross-validation results for the category *Inflammation*. Evaluation results marked with * exhibit a significant difference to the baseline (Snowball) marked with ’ ($p < 0.05$). The maximum value is underpinned in gray.

malization component led to a significant loss in evaluation performance for both tasks. Once chosen a term representation scheme (*Snowball*, *Mixed*, *Morpho*) with or without a character n-gram model, also the weighting scheme tends to have no influence on the classification task anymore. The performance is fixed via the other two parameters.

Table 4.5 shows that the minimal performance difference to the best classification result without using a morphosemantic component, can be reached with a character 4-gram model and a binary weighting scheme. In this category as mentioned before applying other weighting schemes has no noticeable performance impact on the particular classification task and the classification performance is basically the same as using a morphosemantic component.

This comparison can also be applied to Table 4.2 where a minimum performance difference to the best classification result excluding morphosemantic processing is achieved via a 3-gram decomposition with the I(n)B2 weighting scheme. As within this category no significant performance difference according to the chosen weighting scheme is depicted, the one with the lowest processing power can be chosen, which is a binary weighting scheme in this case.

As applied character 4-gram models to the output of the Lucene SnowballAnalyzer makes a positive significant difference within *all* classification results regardless what weighting scheme was used it can be argued that this is the favorite character n-gram decomposition if *no* morphological component is available. If a morphological component is in use a simple binary weighting scheme can be used without character n-gram models, reaching a slightly better performance merging with tokens from the SnowballAnalyzer. The

performance decrease using a Lucene SnowballAnalyzer and a binary token based weighting scheme together with a 4-gram decomposition compared to the maximal performance parametrization is negligible for inflammation- and small for neoplasm detection.

4.5 Conclusion and outlook

In this chapter it was investigated to what extend a parameter optimized machine learning method especially known to have a good performance on textual data can be applied to clinical document classification. For this an already available gold standard of pathology reports was used to test an advanced classification approach, especially the impact of a morphological preprocessing component on the annotation performance was of interest. For inflammation and neoplasm detection two SVMs exploiting a linear kernel were trained for the concrete task. One time the Lucene SnowballAnalyzer and another time morphosemantic normalization, exploiting this component from the Averbis Extraction Platform, where applied to the raw pathology reports. Out of this a third text base was formed merging both obtained tokens into one document. (2-5)-gram models were built in the next step and four different weighting schemes (bin, tf, tf-idf, I(n)B2) applied on the feature vectors obtained from the originated term-document matrix. The different trained SVMs were evaluated using a 10-fold-cross validation.

For neoplasm detection a micro-averaged F-measure of 0.96 was obtained. Inflammation classification reached a micro-averaged F-measure of 0.97. Both classification tasks reached its maximum performance using the the joined tokens from the Lucene SnowballAnalyzer and the morphosemantic normalization. 3-grams together with I(n)B2 were applied to get the maximum performance level for neoplasm detection. For the inflammation classification task a binary weighting scheme without character n-gram models led to maxmimal peformance.

Nevertheless a more detailed look on the results revealed the following facts for this clinical document classification use case:

- The applied weighting scheme has predominately no relevant impact on classification performance under the condition of a chosen character n-gram model and tokenization method.
- 4-gram models always have a positive impact on classification performance if no morphosemantic component is used.
- 2-gram models always result in a negative impact on classification performance if a morphosemantic component is used.
- If no character n-gram model is used, merged tokens from a morphosemantic analysis and a SnowballAnalyzer always have a positive impact on classification performance.

Out of this observation inflammation detection can be fulfilled with a micro-averaged F-measure of 0.97 using a SnowballAnalyzer together with a character 4-gram model and a binary weighting scheme alone. The performance difference to the maximum value using

a mixed token base is insignificant ($\Delta_{F_1} = 0.002$). For neoplasm detection using this less calculation expensive and straight forward scheme, a performance level of 0.95 is reached, having a small but relevant performance drop ($\Delta_{F_1} = 0.012$). From this investigation four simple observations can be deduced to reach high document level classification results with respect to realistic processing constraints when using a supervised machine learning approach. A SVM exploiting a linear kernel can be used for document classification in a clinical environment achieving high F-measure values ≥ 0.95. A binary weighting scheme is sufficient for this task. A 4-gram model should be applied to the output of a Lucene SnowballAnalyzer if no morphosemantic component can be integrated. If a morphosemantic component is available a binary weighting on a mixed approach (SnowballAnalyzer $\cup$ Morphosemantic normalization) without character n-gram decomposition should be applied.

It is quite remarkable that with this advanced bag-of-words approach F-measure values greater than 0.96 are possible. Nevertheless, when comparing this machine learning approach to the results of the regular expression based knowledge-engineered system with an F-measure 0.98 [62], the performance difference is noticeable. Using the machine learning approach, the training of the model is time consuming, but the classification itself is fast when using a SVM. The core mathematical procedure is a sum of multiplications. One time-critical issue is to transform the document to a vector representation that is used as input feature vector for the classifier. The weighting scheme I(n)B2 is time consuming and not practical in contrast to a simple binary weighting scheme.

Two possibilities could therefore be applied in a clinical environment, calculating the feature vector representation on the fly or store this representation additionally to the original text beforehand. A disadvantage of the machine learning approach is its black box style statistical classification model, which makes decisions hard to comprehend. Within a rule based system, the traceability of decisions is clear.

Future work could address other machine learning methods than SVMs. In addition, the character n-gram scheme could be varied, e.g. extended to a token-level. The influence of this variation on the classification performance has to be investigated. The impact of additional annotations from a NLP pipeline made to the original text and to what extend they can be exploited as features in a machine learning approach could be highlighted. E.g. if an enriched version of the original document with SNOMED CT concepts has a positive impact on the classification task. Crucial to this investigation are always (i) system performance in the sense of the quality of the annotations and (ii) real time constraints, as additional processing and modeling steps can influence heavily processing time.

Chapter 5

Clinical Information Retrieval

5.1 Introduction

Despite the emergence of data processing technologies like NoSQL approaches and data structures supporting the processing of big data sets, most current CIS environments are based on classic database management systems. In each jurisdiction they need to address legal requirements, e.g., that documents have to be archived for years. This requires that large pools of legacy data need to be kept accessible. Despite guaranteeing access to growing pools of clinical documents, supporting accounting modalities and the support of information needs resulting from research questions stated by physicians are other IR use cases. For exploiting so called secondary use scenarios, targeted content extraction is therefore increasingly seen as a major necessity for CISs.

The Scientific Service Area - Medical Data Management group at the Institute for Medical Informatics, Statistics and Documentation (IMI) maintains services and tailored database connectors to the hospital information system of the Graz University Hospital, addressing information needs formulated by physicians and researchers. Frameworks that support ETL workflows (e.g. Talend Open Studio or i2b2) have proved useful for providing an integrated access to the backend of CISs by large sets of data base connectors. For fulfilling these requests, wild card based searches within SQL-like queries, access free text fields in database systems. Given the preference of recall-oriented search strategies and the complexity of medical language, such queries are often quite complex as they combine regular expressions with classical database indexes. The quality of the retrieval results therefore depends on the human translation of the physicians' information need into the supported query language, which requires considerable skills and knowledge.

In this chapter, IR scenarios based on clinical information needs are used to evaluate retrieval models that exploit distributional semantics in combination with morphosemantic processing. It is investigated to what extent a statistical indexing scheme can be used with clinical narratives using evaluation methodologies in accordance with the scientific NLP and IR community. A focus is set on handling clinical narratives for search. The chapter is organized as follows: Section 5.2 provides a literature survey regarding recent statistical retrieval approaches applied to EHRs. In Section 5.3 methods, datasets and

frameworks required for evaluation are described. Section 5.4 presents and discusses the evaluation results. Section 5.5 summarizes the work and gives an outlook towards further investigations.

5.2 Related work

Distributional semantics exploiting the VSM and using different dimension reduction methods (LSA, Probabilistic Latent Semantic Analysis (PLSA) [113], Latent Dirichlet Allocation (LDA) [114]) were mainly applied to biomedical proof read text for IR [84, 115–118]. There exist just some studies using these models on clinical narratives for retrieval purposes. The following literature survey gives an overview of distributional semantics with a special focus of LSA applied to the *clinical domain*. It reflects the novelty of the retrieval approach presented in this chapter.

One year after Deerwester et al. published a seminal paper about indexing by LSA [89], Chute et al. [119] applied it to the UMLS Metathesaurus on three different data sets. *Tiny-Input* consisted of 10 concepts of the semantic type "Disease or Syndrome", 101 were assembled for the *Midi-Input* and they used 2,580 concepts for the *Maxi-Input* set. The concepts were applied as documents specifying the row entries, word types of the concept itself together with their synonyms, lexical variants and associated expressions formed the term space with a previous canonicalization of terms using a morph tool [120] (e.g. the terms "tumor", "neoplasm", "carcinoma", and "malignancy" are mapped to the same word type "cancer"). They also created a complex matrix representation, adding an imaginary component, the numbers in there, reflecting the distance of word types out of a hierarchical relation. According to their conclusion, this methodology is promising for retrieving textual patient data.

In the following year, Chute and Yang [121] evaluated concept-based Latent Semantic Indexing (LSI) on surgical case report texts. They applied a two stage similarity score where each column in the matrix represents the cosine similarity to a given concept. 259 concepts were used (35.00 - 39.99 ICD-9-CM procedure codes). Fifteen information needs were defined together with their corresponding ICD-9-CM code and relevant cases according to the information needs marked by two experts. Human based ICD-9-CM code based search performed best, followed by the SMART retrieval system. A simple VSM for final cosine similarity performed better than then the LSI approach. Dimensionality reduction had little influence on the evaluation results, in contrast to which parts of the document were used to generate the code mappings. Using the procedure names only performed better than the surgical report only, or both fields combined. The poor performance was attributed to sub-optimal initial concept matching due to the fact that synonyms as well as concept similarity vectors were used for the final ranking of the retrieval results. A more complex weighting scheme instead of the binary one could improve the results further. Finally, the authors argued that LSI and retrieval is based on document to document similarity without a concept layer in-between as exploited in their two stage similarity score.

Kintsch [122] introduced the idea of LSA in the context of mashine grading of clinical case summaries. Firstly, an appropriate semantic space had to be constituted using medical dictionaries, textbooks and a certain amount of sample case summaries. One strategy for a proper grading assumes a set of pre-graded essays, represented as a cluster in the semantic space. A newly written essay is mapped into this space, and the automatically assigned grade is a combination of weights with respect to their cosine similarity to the 10 nearest pre-graded essays. Another strategy would be to define a set of gold standard essays written by experienced medical experts, and a grade of a new essay is a function of its cosine similarity to the expert essays.

Cohen and Hersh [123] applied LSA to psychiatric treatment. The idea was to assign dangerousness scores to psychiatric narrative reports according to their content. Two use cases were selected: "dangerousness to self" and "dangerousness to others". Within these two classes scores were assigned from 1 to 5. From 51,524 documents harvested from different sources a 100 dimensional semantic vector through LSA was built for each document. The 392 psychiatric summaries, rated before to the aforementioned dangerousness scores and categories were also mapped into this space. Machine calculation of a score was done by weighting the 20 nearest summaries according to their distance in the semantic space to the summary under scrutiny mapped into the same semantic space. A Spearman test between human- and computer- assigned dangerousness grades resulted in $r_s = 0.41$ for "danger to self" and $r_s = 0.55$ for "danger to others" ($p < 0.0001$). The authors suggested a combination of NLP and LSA to address negation and vagueness, as this was not taken into account in their bag-of-words model underlying LSA.

Cohen et al. [124] used LSA to harvest clinical concepts from a big corpus of psychiatric narrative reports, in order to cover the possible type space which was necessary for processing the psychiatric narratives. They reduced the term-document matrix using the General Text Parser Tool (GTP) and reduced its dimension to 303 for 50,028 documents. Five so-called facet models were selected for training and evaluation of the system, including Psychosis, Mood Disorder, Substance Abuse, Violence and Suicide. For each facet, derived terms and phrases were extracted manually according to the Diagnostic and Statistical Manual of Mental Disorders - Revision Four (DSM-IV) criteria [125]. Out of this, a facet model representing a subspace [126] of the initial semantic space was defined. In a second step, the semantic vectors defining the sub space were trained and optimized using 100 narratives split up into propositional units. Each unit was manually assigned to a facet model. System-rater recall was 0.77, precision 0.97 and positive specific agreement 0.71. Rater-rater evaluation achieved a recall of 0.77, precision of 0.99 and positive specific agreement of 0.81. This implies the hypothesis that machine comprehension of clinical facts can be implied to achieve as good results as humans doing the same task.

Ginter et al. [127] combined LSA with Hidden Markov Models (HMMs) in order to automatically perform the task of topic segmentation in an unsupervised way. A pool of intensive care nursing narratives from 135 patients was used, which were divided into a training (198 nursing shifts) and test data set (204 nursing shifts). Different passages in these reports were assigned to certain topics e.g. hemodynamics, diuresis, breathing. As expected, the supervised HMM based approach outperformed the unsupervised model with an accuracy of 0.83 vs. 0.75. The supervised method needed only about 3,600 la-

beled words to reach the performance of the unsupervised method, which reaches its peak performance using about 360,000 unlabeled words.

Tremblay et al. [128] tried to identify fall-related injuries from EHRs. They built a term-document matrix out of text snippets and applied two different weighting schemes on the matrix, *viz.* an entropy-based weighting scheme on the one hand and a scheme based on information gain on the other hand, which took into account whether a document is about a fall-related injury or not out of a gold standard with 2,157 care episodes. After applying the matrix weighting a SVD was performed but no dimension reduction. The documents were clustered in this space first using an unsupervised approach using k-means, and then using a supervised method exploiting logistic regression. The combination of information gain weighting together with logistic regression performed best, achieving an accuracy of 0.91 on the test data set (20% split; 434 care episodes).

Elvevåg et al. [129] applied LSA in order to objectively analyze formal thought disorder, an anomaly typical for schizophrenic patients. They build an LSA-based semantic space reflecting the general reading ability of a healthy person of a first year university level. The corpus consisted of 37,651 text samples and 92,408 word types. Applying LSA the number of dimensions was reduced to 300. A structured interview was applied to two groups, with one group fulfilling the DSM-IV criteria for schizophrenia against a healthy control group. Four basic measurements were conducted using the semantic space, based on the structured interview: relatedness between words, relatedness between sentences, relatedness of an answer to a question and how answers of different persons to a questions relate to each other. The authors could show how LSA can be used as theoretical background for an instrument that allows schizophrenia diagnosis based on detecting formal thought disorder.

For the sake of completeness of the literature review on distributional semantics, enhancing the scope of applied LSA in the clinical domain, the following publications should be mentioned: Henriksson et al. investigated in detail how random indexing methods can be applied in three different areas: diagnosis coding support [130–133], synonym (and abbreviation/expansion) extraction [134–137], adverse drug reaction - event exploration/detection [138] and NER [139]. Distributional semantics are also intensively applied in the field of topic modeling using the unsupervised LDA [8–10]. Recent work applying distributional semantics within clinical retrieval and extraction scenarios are demonstrated by Moen et al. [140], Zhang and Elhadad [141] and Natarajan [142]. A good overview about distributional semantics in general can be found in Cohen and Widdows [143].

In contrast to the methodological approaches in this review, the work presented in this chapter follows a strict TREC-based evaluation guideline, described in more detail in Section 5.3.2. With a fixed number of minimal information needs for statistical IR evaluation, it is investigated how different model parameters influence the overall retrieval performance. This semantic normalization design space of the retrieval model is explained in Section 5.3.1. The approach is used to estimate to what extent the degree of dimension reduction hampers the IR model, and to find the optimal number of features for the retrieval model in this use case. To the best of the author's knowledge this is the first approach of this kind applied to German clinical narratives.

5.3 Materials and methods

5.3.1 Overview

In order to investigate the general applicability of distributional semantics for retrieval purposes on clinical narratives we define our experiments on a semantic normalization design space for the retrieval model, containing seven axes. The first axis describes the dimension reduction methods used, in our case LSA (Section 3.3). The second axis is the degree of dimension reduction. Too aggressive as well as too moderate dimensionality reduction hampers retrieval performance [144]. The third axis defines the weighting scheme of the VSM before dimensionality reduction. Four different weighting schemes are applied: binary, term frequency, term frequency-inverse document frequency and I(n)B2, explained in Section 3.2.1. The fourth axis defines the degree of textual preprocessing or annotation steps. We applied one time the straight forward Lucene SnowballAnalyzer for tokenization, another time morphosemantic normalization for obtaining the index tokens, described in detail in Section 3.5 and finally a combined approach. Examples are provided in Section 4.3.3. The sixth axis defines the similarity measures used in the Euclidean space. The cosine similarity is applied in this case (Section 3.2.2). The last axis defines whether query expansion methods are applied or not. In this approach no automatic expansion methods are applied. A defined setting of this model is then applied to a certain language (German) and (sub-)domain (medicine, pathology). This investigation is made on an extended version of the gold standard described in Section 4.3.2 and comprises about 3500 pseudonymized pathology reports. MAP and Precision at k ($P_{10}, P_{20}, P_{30},$) are used as evaluation measures for ranked retrieval results (Section 3.6) per information need and parametrized semantic design space as described above.

5.3.2 Information Needs

According to TREC-based evaluation guidelines [14] and based on the amount of defined information needs of the TREC Medical Records Track [145, 146], 26 information needs were defined for retrieval evaluation. The information needs had been chosen by an analysis of pathology based retrospective requests from physicians to the Scientific Service Area - Medical Data Management group at the Medical University of Graz. One criterion of choosing the information needs was that a least 30 relevant documents could be extracted from the existing pathology report gold standard described in Section 4.3.2. Table 5.1 gives an overview of the defined information needs.

We understand by "information need" a supposed cognitive representation in the mind of a person who needs to fill an information gap 2.3.2. In a search process using a search engine this person verbalizes this information need using their natural language (e.g. German, English, French) and language register (e.g. medical expert, medical student, layperson). In the existing database with 26 queries, containing links between queries with relevant documents, the information needs had been formulated inspired by real expert queries constructed by medical specialists who intended to define cohorts from their patients according to clinical features. In this new experiment, we rephrased these queries in a standardized way. As a reference, we attempted to reconstruct each information need by

Topic number	Information need	Inflammation	Neoplasm	Dignity	Location	#Relevants
1	Gastritis	•	-	-	Stomach	567
2	Hepatitis	•	-	-	Liver	73
3	Colitis	•	-	-	Colon	140
4	Appendicitis	•	-	-	Appendix	196
5	Dermatitis	•	-	-	Skin	53
6	Cholezystitis	•	-	-	Gallbladder	83
7	Cervicitis	•	-	-	Cervix	62
8	Duodenitis	•	-	-	Duodenum	41
9	Prostatitis	•	-	-	Prostate	51
10	Sinusitis	•	-	-	Paranasal sinuses	44
11	Adenkarzinom in the colon	-	•	malignant	Colon	146
12	Neoplasm in the intestine	-	•	benign, malignant	Intestine	374
13	Neoplasm in the prostate	-	•	benign, malignant	Prostate	88
14	Breast cancer	-	•	malignant	Breast	66
15	Adenom in the intestine	-	•	benign	Intestine	167
16	Nevus	-	•	benign	Skin	56
17	Adenom	-	•	benign	-	211
18	Lipom	-	•	benign	-	35
19	Myom	-	•	benign	-	50
20	Malignom	-	•	malignant	-	674
21	Carcinoma	-	•	malignant	-	487
22	Adenocarcinoma	-	•	malignant	-	325
23	Neoplasm	-	•	benign, malignant	-	1125
24	Inflammation	•	-	-	-	1657
25	Derma	-	-	-	Skin	233
26	Breast	-	-	-	Breast	130

Table 5.1: Defined information needs for evaluation purposes.

using headings and text words in the German language web-based health consumer portal www.netdoktor.at.

The term preferences in *Netdoktor* are somewhat in between layperson and expert terms, according to the fact that in Germany and Austria many expert terms, like "Diabetes" or "Prostata" are also understood by patients, although local terms exist. We varied the *Netdoktor*-inspired queries in two ways. In the first variation we tried to find a typical layperson expression for each information need, in the second one a typical expert expression. Every information need was transformed into a query, according to the query syntax supported by the IR tool and with respect to the language register (*Layperson*, *Netdoktor*, *Expert*). In this evaluation, the information need is transformed to a set of expressive terms describing the content of the request. The transformed information needs with their terms are depicted in Table 5.2.

5.3.3 Gold standard

The relevant documents for the information needs have been extracted via exploiting the annotations of the gold standard described in Section 4.3.2. A schematic inflammation-location SQL query is analyzed and explained in the following:

```
SELECT ID,DI FROM befunde_gs1 JOIN details_gs1 ON befunde_gs1.ID =
    details_gs1.BefundID WHERE ((details_gs1.I_Organ_Text Like '53%')
    AND (details_gs1.I_Sicherheit < 8)
```

	Language register		
Topic number	Layperson	Netdoktor	Expert
1	Magenschleimhautentzündung	Gastritis	Gastritis
2	Leberentzündung	Hepatitis	Hepatitis
3	Entzündung des Dickdarms	Entzündung des Dickdarms	Kolitis
4	Blinddarmentzündung	Blinddarmentzündung	Appendizitis
5	Entzündung der Haut	Dermatitis	Dermatitis
6	Gallenblasenentzündung	Gallenblasenentzündung	Cholezystitis
7	Entzündung des Gebärmutterhalses	Entzündung des Gebärmutterhalses	Zervizitis
8	Entzündung des Zwölffingerdarms	Entzündung des Zwölffingerdarms	Duodenitis
9	Entzündung der Vorsteherdrüse	Prostatitis	Prostatitis
10	Nebenhöhlenentzündung	Nebenhöhlenentzündung	Sinusitis
11	Adenokarzinom des Dickdarms	Adenokarzinom des Dickdarms	Adenokarzinom des Kolon
12	Geschwulsterkrankung des Darms	Tumorerkrankung des Darms	intestinale Neoplasie
13	Geschwulsterkrankung der Vorsteherdrüse	Tumorerkrankung der Prostata	Prostatatumor
14	Brustkrebs	Brustkrebs	Mammakarzinom
15	Adenom des Dickdarms	Adenom des Dickdarms	Adenom des Kolons
16	Muttermal	Muttermal	Nävus
17	Adenom	Adenom	Adenom
18	Geschwulst des Fettgewebes	Lipom	Lipom
19	Geschwulst des Muskelgewebes	Myom	Myom
20	Bösartige Geschwulst	Bösartiger Tumor	Malignom
21	Krebs	Krebs	Karzinom
22	Bösartige Geschwulst aus Schleimhaut oder Drüsengewebe	Adenokarzinom	Adenokarzinom
23	Geschwulsterkrankung	Tumorerkrankung	Neoplasie
24	Entzündung	Entzündung	Entzündung
25	Haut	Haut	Dermis
26	Weibliche Brust	Brust	Mamma

Table 5.2: Query terms according to the language register in use.

Important are the SQL syntax parts `(details_gs1.I_Organ_Text Like '53%')` and `(details_gs1.I_Sicherheit < 8)`. `53%` specifies all body parts belonging to the ICD-O-3 topographical code C53 together with its subcodes, therefore in this example matching "C53: Cervix uteri", "C53.0: Endocervix", "C53.1: Exocervix", "C53.8: Overlapping lesion of cervix uteri", "C53.9: Cervix uteri". The second part includes all inflammation specific annotations making a statement about the confidence of the existent of an inflammation in this case. Following cases are included: 1 $\widehat{=}$ "probably", 2 $\widehat{=}$ "compatible with", 3 $\widehat{=}$ "questionable", 4 $\widehat{=}$ "possible", 5 $\widehat{=}$" uncertain", 6 $\widehat{=}$ "unlikely", 7 $\widehat{=}$ "safe" except case 8 $\widehat{=}$ "certainly not" and case 9 $\widehat{=}$ "not mentioned". This reflects the fact of a *recall oriented* gold standard for the specific information needs.

The second schematic SQL query for the identification of relevant documents of neoplasm-location like information needs looks like:

```
SELECT ID,DI FROM morphologie RIGHT JOIN details_gs1 ON morphologie.SUI
    =details_gs1.K_Morphologie JOIN befunde_gs1 ON befunde_gs1.ID =
    details_gs1.BefundID WHERE

(((details_gs1.K_Organ_Text) Like '18%') AND ((details_gs1.K_Typ)>0
    And (details_gs1.K_Typ)<10) AND ((morphologie.Bezeichnung) Like '%
    adeno_ar_inom%') AND ((details_gs1.K_Sicherheit)<8))

OR
```

```
(((details_gs1.K_Organ_Text) Like '19%') AND ((details_gs1.K_Typ)>0
    And (details_gs1.K_Typ)<10) AND ((morphologie.Bezeichnung) Like '%
    adeno_ar_inom%') AND ((details_gs1.K_Sicherheit)<8))

OR

(((details_gs1.K_Organ_Text) Like '20%') AND ((details_gs1.K_Typ)>0
    And (details_gs1.K_Typ)<10) AND ((morphologie.Bezeichnung) Like '%
    adeno_ar_inom%') AND ((details_gs1.K_Sicherheit)<8))
```

`(details_gs1.I_Organ_Text Like '18%')` again refers to ICD-O-3 topographical annotations starting with C18, and its subcodes. In this example it is refering to "C18: Colon", "C18.0: Cecum", "C18.1: Appendix", "C18.2: Ascensing colon", "C18.3: Hepatic flexure of colon", "C18.4: Transverse colon", "C18.5: Splenic flexure of colon", "C18.6: Descending colon", "C18.7: Sigmoid colon", "C18.8: Overlapping lesions of colon" and "C18.9: Colon, NOS". Annotations of the type `K_Typ` express the dignity of the neoplasm again referring to ICD-O-3 and the standardized description for these codes. In this example all malignant ones were searched for (morphology code, behavior digit: 1 $\widehat{=}$ "uncertain behaviour", 2 $\widehat{=}$ "carcinoma in situ", 3 $\widehat{=}$ "malignant, primary site", 6 $\widehat{=}$ "malignant, metastatic site", 9 $\widehat{=}$ "malignant, uncertain whether primary or metastatic site") and the benign ones were left out (morphology code, behavior digit: 0 $\widehat{=}$ "benign"). The annotation type `Bezeichnung` has its value set and the corresponding description of the morphological ICD-O-3 codes. Just those annotations are returned which match the wildcard based query `%adeno_ar_inom%`. This wild card based search also reflects the fact of the ambiguity of the use of the characters "c", "k", "z" in the medical language, explained in more detail in Section 2.2. The search expression repeats itself in this case for three times with alternating topographical search statements, building the union of the result sets at the end. Finally the overall SQL search expression returns all relevant documents for the information need "Adenokarzinom in the colon", one of the defined information needs listed in Table 5.1.

5.4 Results and discussion

Table 5.3 shows the overall evaluation of the retrieval approach using LSA depending on the various axes defined in the evaluation model. The table depicts the maximum MAP for a certain evaluation setting, reflecting the fact that we get a maximum performance of our retrieval model using the language register of *Netdoktor*, together with the weighting scheme I(n)B2, applying morphosemantic analysis, having its maximum performance at a feature space dimension of 15, reaching a MAP of 0.55.

Having the local MAP maxima with respect to the applied degree of dimensionality reduction at a level of 15 indicates high performance with *aggressive feature space reduction*. Figure 5.1 shows the changing MAP value with respect to the number of features in use for the weighting scheme I(n)B2. All three examined language registers depict a MAP maximum value using morphosemantic analysis as preprocessing method. Interestingly

	bin	tf	tf-idf	I(n)B2
Layperson				
Snowball	0.13_{35}	0.12_{15}	0.12_{120}	0.13_{25}
Morpho	0.34_{55}	0.36_{75}	0.36_{110}	0.51_{15}
Mixed	0.33_{55}	0.34_{100}	0.36_{150}	0.48_{20}
Netdoktor				
Snowball	0.25_{35}	0.23_{55}	0.22_{245}	0.26_{55}
Morpho	0.37_{55}	0.38_{75}	0.39_{145}	0.55_{15}
Mixed	0.37_{55}	0.37_{100}	0.39_{165}	0.51_{20}
Expert				
Snowball	0.37_{190}	0.35_{115}	0.33_{140}	0.41_{55}
Morpho	0.39_{120}	0.41_{75}	0.41_{95}	0.54_{15}
Mixed	0.36_{55}	0.39_{120}	0.40_{130}	0.51_{15}

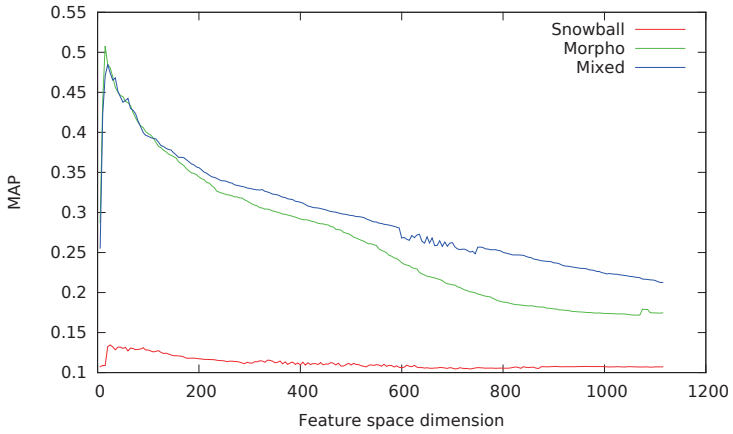
Table 5.3: Local maximum of the MAP with respect to the degree of dimensionality reduction, weighting scheme, language register and preprocessing methodology. The maximum value is underpinned in gray.

the token combination of morphosemantic analysis and a Lucene SnowballAnalyzer beats pure morphosemantic analysis with increasing feature space.

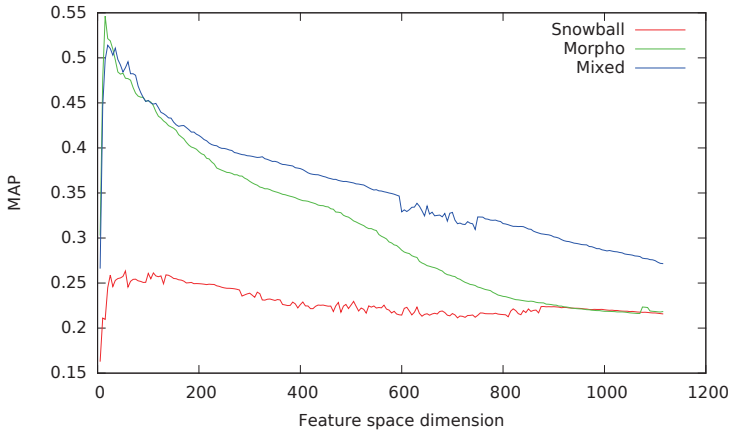
Having the highest MAP performance at a relatively small feature space dimension can be interpreted in a way that just with a few perpendicular axes which build the semantic space, the information content can be properly described with respect to the information needs used for evaluation. Adding more axes to the semantic space decreases retrieval performance, which can be seen as modeling noise into the retrieval model. It is quite remarkable if the recommended feature space dimensionality from literature would have been applied to LSA, according to Bradford [144] $k \approx 400$, a significant performance decrease compared to our actual maximum at a very low semantic space would be the effect. This observation is true for all three language registers (Figure 5.1).

The weighting schemes for building the term-document matrix also have a remarkable impact on the overall performance of the retrieval model, cf. Figure 5.3. In accordance to the literature, I(n)B2 has the highest positive impact on retrieval performance compared to the other weighting schemes, achieving a maximum performance gain of nearly 49% compared to a naive binary weighting scheme approach. In addition, the model with the highest retrieval performance used the I(n)B2 weighting scheme. The term frequency and term frequency-inverse document frequency weighting schemes had a much lower impact on retrieval performance compared to binary weights.

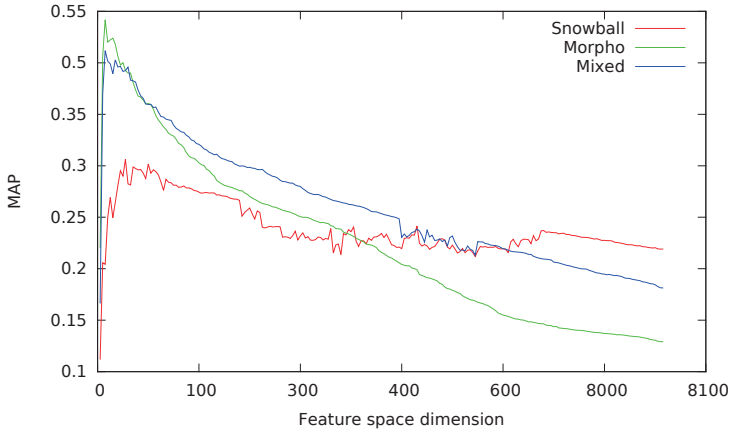
The impact of the language register, to express the information need under test, has especially an influence in case just a SnowballAnalyzer is used to generate the dictionary for the term-document matrix. In this case, there is a performance gain of up to 100% when switching from the language register *Layperson* to *Netdoktor*, and an increase of about 58% when using the *Expert* language register instead of *Netdoktor*. This performance boost



(a) Language register: *Layperson*



(b) Language register: *Netdoktor*



(c) Language register: *Expert*

Figure 5.1: Retrieval performance according to different degrees of dimension reduction for different language register in use and the weighting scheme I(n)B2.

can mostly be explained by the fact that *Netdoktor* and *Expert* are the language registers that best match the terms in the document corpus.

The overall performance of using a SnowballAnalyzer alone reaches its maximum value in combination with the I(n)B2 weighting scheme and the language register *Expert*, at a MAP level of 0.41. The maximum language register dependent performance difference in this case is 0.28, compared to the language register *Layperson*. This performance difference decreases significantly when morphosemantic processing or a mixed approach is used before building the term document matrix for the retrieval model under test. The performance difference due to the use of a different language register drops from a maximum of 0.28 to 0.04. Morphosemantic preprocessing therefore supports the results with non-expert language registers up to the point that similar retrieval results are obtained compared to the language register *Expert*. The different initial dictionary sizes of the term-document are shown depicted in Figure 5.2. Applying the Lucene SnowballAnalyzer results in a feature space dimension respectively dictionary size of 1,936. Morphosemantic indexing reduces the feature space of about a half to 1,115 word types. The combined variant has the biggest initial feature space dimension of 3,050.

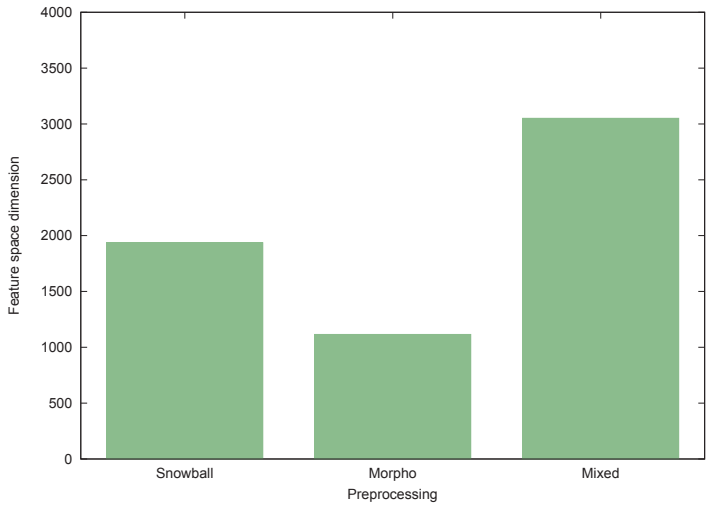


Figure 5.2: Feature space dimension of the initial term document matrix with respect to its preprocessing method.

The positive effect of the morphosemantic component is particularly demonstrated with the performance values for topic #16 in Appendix A.2.1 Table A.1. This result table contains the evaluation of the IR models using the language register *Layperson* for expressing the information need: "Return all observation reports that are about a nevus". With a very poor performance without the morphosemantic processing step at the beginning, $\text{Snowball}_{25}^{InB2}$: MAP=0.02, $P_{10} = 0, P_{20} = 0, P_{30} = 0$, the results tremendously increase including this component, $\text{Morpho}_{15}^{InB2}$: MAP=0.98, $P_{10} = 1, P_{20} = 1, P_{30} = 1$; Mixed_{20}^{InB2}:

MAP=0.98, $P_{10} = 1, P_{20} = 1, P_{30} = 1$. The evaluation values for this information need regarding the language register *Expert* are shown in Appendix A.2.1 Table A.3. Here, using the SnowballAnalyzer alone good performance results are achieved. This implies that the *Expert* language register search terms were also used in the document collection. The performance is further improved by use of the morphosemantic component, resolving synonyms. As can be seen in the following example the search term "Muttermal", used in the case of the *Layperson* and *Netdoktor* language registers and "Nävus" the *Expert* search term, are mapped to the same representation after morphological normalization *viz.* $\text{Muttermal}_{Layperson}$: nevoidiikrxw; $\text{Muttermal}_{Netdoktor}$: nevoidiikrxw; Nävus_{Expert}: nevoidiikrxw.

Analyzing the information need of topic #8, "Return all observation reports which are about duodenitis", also by exploiting the *Expert* language register together with morphosemantic normalization and I(n)B2 weighting a low performance is achieved. This can be explained by the fact that in this case the dimension reduction is too aggressive for this particular information need. Nevertheless, when averaging over all information needs the estimated number of dimensions (n=15) for the SVD-document matrix performs best of all possible dimensions. The maximum performance for this particular information need is reached at $\text{Morpho}_{190}^{InB2}$: MAP=0.37, $P_{10} = 0.5, P_{20} = 0.5, P_{30} = 0.47$. This observation again underlines the difficulty of a correct dimension size estimation for the IR model.

5.5 Conclusion and outlook

This chapter focused on the applicability of a retrieval approach exploiting distributional semantics on a classical document retrieval task. For this purpose, a design space model for the IR approach under test was defined and evaluated, and the resulting retrieval models were evaluated according to this design space. Our best retrieval model achieved a performance of $\text{Morpho}_{15}^{InB2}$: MAP=0.55 exploiting the *Netdoktor* language register. It could be shown that aggressive dimension reduction in this case led to the best overall retrieval performance. The performance of the IR models noticeably decreases with higher level of dimensions. The performance difference of the evaluated maximum MAP, using 15 features of the best retrieval model and the same retrieval model using the degree of dimensions given in literature ($\text{Morpho}_{400}^{InB2}$: MAP=0.34) is 0.21. This suggests that optimizing the model with respect to the features in use should be done for every model built for a certain domain.

The morphosemantic normalization component has a clear positive performance impact in the sense that the use of *non-expert language* exploited for translating the information need to the expression syntax of the retrieval tool achieves similar performance than using *expert language*. In contrast to the positive impact on retrieval performance, morphosemantic normalization is a complex mechanism in the background with an impact on processing time.

Beside this positive impact of the morphosemantic component, the mathematical model of LSA has the advantage that it supports the recognition of synonyms and relates similar concepts. On its downside is its high computational complexity because of the SVD. In

these experiments, computing three SVDs (*Snowball*, *Morpho*, *Mixed*) took about 190 minutes with a maximum matrix dimension size of 3050 times 3542 on a multi-threaded environment (Intel(R) Core(TM) i7-3770 CPU@3.40 GHz 32GB RAM 64Bit-Win7 OS). For the use in CISs with millions of documents, frameworks for parallel processing of the SVD have to be exploited e.g. with the Apache Mahout package or considering methods described by [147–150]. Another possibility is the use of random indexing methods [151] in order to tame computational complexity.

In a clinical environment it also has to be considered that the pool of documents is dynamic and the indices have to be dynamically updated. This requirement is especially challenging with this model as the weighting of the initial term document matrix has to be recalculated. Therefore it is arguable that this method is more suitable for static document pools e.g. a set of legacy data.

The fact that these models treat the documents as bag of words also has to be considered in a retrieval scenario applied to clinical environments. Language grammar or e.g. negations are not considered in a bag of words approach. Clinical narratives very often contain negations in different variations depending on the local language and documentation school [152]. Semantics between terms are based on co-occurrence statistics over the whole document collection and not through syntactical analysis of the grammar structure. Another desideratum is the fact, that clinical documents extracted of the CIS are typically semi-structured containing content specified with a certain header. Typically e.g. is the family history containing facts not directly connected to the patient. Such meta-data information about content should be taken into account in an clinical IR system and a loss of such structured meta-information via e.g. scanning and archiving clinical documents hampers advanced retrieval approaches.

Future work can consider applying other dimension reduction methods like PLSA or LDA and its impact in combination with morphosemantic processing on retrieval performance when using clinical narratives. According to the literature both methods should have a positive performance impact on the distributional semantics approach when using proof read newspaper or biomedical texts e.g. abstracts from scientific data bases. Also the impact of automated query expansion exploiting semantic relations within an ontology could further be evaluated for high level queries e.g. using umbrella terms. This leads directly to a more concept then a term based approach as the query and therefore the document pool have to be mapped to the concepts of the semantic data structure which should be exploited. Applying SNOMED CT concept mapping on (i) a term level approach and (ii) on morphosemantic stems should further on have a positive impact on recall oriented search strategies.

Chapter 6

Clinical Information Extraction

6.1 Introduction

Cohort studies in a clinical setting typically aggregate a wide range of clinical data, for which different subsystems within a CIS need to be accessed, characteristically resulting in a mix of structured and unstructured data. In addition, there is increasing interest in connecting sample data from biobanks with clinical data, in order to gain insight into geno-phenotype dependencies. Whereas biobanks [153, 154] typically store population-based and disease-focused collections of biological materials, patient-related information is stored in heterogeneous and modular CISs. It is investigated to what extent clinical texts constitute appropriate sources from which structured information for clinical research can be reliably extracted. Emphasis is placed on multidisciplinary cohort studies.

The work is organized as follows: Section 6.2 provides a literature survey regarding recent text mining approaches applied to EHRs for disease cohort building. In Section 6.3, methods, data sets and frameworks required for this investigation are described. Section 6.4 presents and discusses the evaluation results. Section 6.5 summarizes the work and an gives an outlook towards further investigations.

This chapter is based on Kreuzthaler et al. [155] following the copy right statement stated in Appendix B.1. The work was supported by the GEN-AU III grant (GATiB II, Workpackage 4) from the Austrian Ministry of Education, Science and Culture.

6.2 Related work

This chapter describes recent work in the area of mining medical records that focus on cohort building based on data extracted from clinical narratives among others, that combine clinical data with data stored in biobank systems. In a cohort study (longitudinal study), the association between exposure and disease is investigated by following individuals (the "cohort") through a time span and measuring the rate of occurrence of new cases in the different exposure groups. A typical example of this kind of study recently started in Sweden with the aim of collecting comprehensive data on lifestyle factors, together with blood

samples, purified DNA and mammograms [156]. The survey starts with examples from medical text mining challenges and then addresses systems and prototypes more related to the use case investigated in this chapter.

Concept extraction, assertion classification, and relation classification applied to clinical text were tasks in the 2010 i2b2/VA challenge [157]. The data set that had been released to the participants comprised 394 training reports, 477 test reports and 877 non-annotated reports. The conditional random fields technique was found to be the most effective method for concept classification, achieving an F-measure of up to 0.92. The most effective assertion classification and relation extraction systems used SVMs as their core methodology, obtaining an F-measure of 0.94 for assertion classification and 0.74 for relation extraction, respectively. Machine learning methods were intensively applied to the specified problems as reported in more detail by Bruijn et al. [158]. Generally, rule based methods were used as supportive instruments for machine learning methods (pre- and post-processing of data).

A further i2b2 challenge was the identification of patient smoker status ("past smoker", "current smoker", "smoker", "non-smoker", "unknown") from medical discharge records [53]. 502 de-identified discharge summaries were used for the challenge. The majority of the systems applied machine learning methods, giving a micro-averaged F-measure of over 0.84. Rule-based methods were mostly applied in combination with classifiers.

Heintzelman et al. [159] applied a rule-based NLP system (ClinREAD), to categorize the pain status in patients with metastatic prostate cancer into four different groups ("no pain", "some pain", "controlled pain", "severe pain"), together with a longitudinal analysis and visualization of the mined status. They applied their system to a patient cohort of 33 subjects forming a text pool of about 24,000 pages and achieved an F-measure of 0.95 for pain detection and 0.81 for pain severity management. The system was further evaluated on the i2b2 corpus showing its generalizability.

A recent approach was described by Skeppstedt et al. [160] concentrating on the extraction of mentions of disorders, findings, pharmaceutical drugs and body structures. They assembled and annotated a corpus of 1,148 randomly selected assessment fields. The selected content represented typical idiosyncrasies of clinical narratives such as telegraphic language, many abbreviations, and few full sentences. Inter-annotator agreement varied between an F-measure of 0.66 (finding) and 0.90 (pharmaceutical drug). The final conditional random field model achieved an F-measure of 0.81 for disorders, 0.69 for findings, 0.88 for pharmaceutical drugs, 0.85 for body structures and 0.78 for the combination of disorders and findings.

Botsis et al. [161] addressed problems with *secondary use of routine data* for a retrospective creation of a pancreatic cancer cohort. They used the data warehouse of the Columbia University Medical Center encompassing 2.7 million patients from which all patient data coded with a "malignant neoplasm of pancreas" and descendants (ICD-9-CM 157.0-157.9) were extracted for a period of 10 years (01/01/1999-01/30/2009). For each patient pathology reports, lab tests, clinical notes and discharge summaries were extracted. Out of the 3,068 identified patients, 1,479 had to be excluded, as no evidence for pancreatic cancer was found in the pathology reports. From the remaining patients, a further 1,067 were excluded due to missing information on core study variables, resulting in a total

of 522 remaining samples. The authors discussed the three most common data quality indicators, *viz.* incompleteness, inconsistency, and inaccuracy. They highlighted *incompleteness* and *inconsistency* as major weaknesses in clinical routine documentation, and as a main problem for the automatic extraction of relevant study information. Another issue identified was the poor assignment of time stamps to the related events, as well as bad documentation quality in the EHRs themselves.

Antolík [162] tested a system for automatic generation of regular expressions to transform the content of clinical narratives into a structured template. They used Amilcare [163], an algorithm for generating regular expressions. An annotated corpus with lemmatization and POS tags processed by an NLP pipeline was used for training. The corpus had a size of 300 documents with a total of 100 different clinical concepts. 40% of the concepts had a frequency greater than 10 in the training corpus. The recognition rate of the more common concepts could be divided into 2 groups: one group had an F-measure > 0.5 and one group an F-measure < 0.1.

Roque et al. [164] analyzed 5,543 EHRs collected over 10 years in a psychiatry department and extended the existing ICD-10 codes with codes resulting from an automatic analysis of free-text content. The tagging approach was tested using records of 48 patients and achieved a precision of 0.88. A negation detection module based on NegEX [152] identified 73% of all relevant negations. The automatically assigned ICD codes were used to analyze comorbidities, to create an ICD-10 disease-based network, and to discover genotype-sided relationships (OMIM [165] was used as a catalog of human genes and genetic disorders). A new genotypic association between alopecia and migraine was shown as a result of this text mining approach.

The Pygargus eXtraction Customized Program, used for building a register of type 2 diabetes mellitus patients, was evaluated by Martinell et al. [166]. One of the following criteria had to be met to identify a patient with type 2 diabetes mellitus: ICD-9 and ICD-10 disease codes, oral antidiabetic drug codes or lab indicators according to the WHO classification of the disease. The system works on both structured and unstructured data, feeding a retrieval system for patient data. The system was tested on a pool of 10,753 EHRs (1993-2005) and compared with a built-in search tool within their CIS. The specificity (true negative rate) was 100%, the sensitivity 99.9% (true positive rate). The authors concluded that automated data extraction can provide a *high coverage* regarding a given disease but it also suffers from a high number of *missing values*.

Xu et al. [167] developed a system for the automatic detection of colon cancer patients, merging structured and unstructured data from EHRs. Using an ETL workflow, data from patients that met certain criteria (ICD-9 code, current procedural terminology code, colorectal cancer keywords, drugs) were extracted from a clinical documentation system from a 10 year period (1999-2008). Of these 17,125 patients, 300 were selected and a gold standard was created. The first step was to find positive colon cancer concepts (document-level concept identification). In the second step, the system was used to detect whether a patient had colon cancer or not (patient-level case determination). Rule-based and machine-learning based methods were combined. The system achieved an F-measure of 0.97 on document-level concept identification, and an F-measure of 0.93 on patient-level case determination.

Segagni et al. [168] describe ONCO-I2b2, a project that combines data from a biobank information system with data from EHRs. They used the i2b2 [34, 169] integration framework, connecting pathology data, sample data and data from their CIS regarding cancer patients. Noun phrases were detected in text passages and mapped to SNOMED CT concepts using GATE [170, 171], a NLP engine. A set of regular expressions was developed for attribute extraction, e.g. one to extract the scoring of a mamma carcinoma. The data was integrated using a complex ETL workflow and is accessible to the end user via a web client. The system has administered 2,214 patients, 25,826 visits, 163 concepts and 93,680 observations so far.

This chapter focuses on the extraction of study-relevant attributes from EHRs. Regular expressions are used in this initial attempt, because the manual tagging of data is expensive and time-consuming. Therefore, in this study a rule-based method is explored, which requires less resources, rather than annotating a training corpus for supervised machine learning. The performance of this rule-based system is evaluated. The documentation quality will be assessed, in particular its influence on the IE approach, from which consequences for mining EHRs in general are discussed. This is crucial, as building cohorts from EHRs is an important use case for the secondary use of EHRs. Also, the preprocessing of data is essential, especially for the integration of clinical data with biobank data, for setting up a geno- phenotype cohort according to a study hypothesis.

6.3 Material and methods

This section details the data set of the investigation and the setup of the evaluation pipeline. Implementation aspects of the framework used for the IE approach are highlighted at the end.

6.3.1 Patient corpus

The main goal of the cohort study, from which a patient corpus was formed, is interdisciplinary research on obesity and related diseases, as well as the comparison of *clinical* and *genetic data*. Patients scheduled for weight-loss surgery underwent a full hepatological and metabolic work-up including liver ultrasound, ultrasound elastometry (fibroscan), clinical and laboratory examination as well as a screening for autoimmune and infectious liver diseases. A lifestyle questionnaire served as a supportive instrument. Blood and tissue samples were acquired during weight-loss surgery and stored at the biobank of the Medical University of Graz. Therefore, the data set constitutes the following three data sources per patient:

- Clinical data:
  - Basic claims data.
  - Medical Clinic Gastro-Enterology (MCGE) reports.
  - Diabetes reports.

  - LIS reports.

- Sample data.
- Lifestyle questionnaires.

Clinical data, Subsystems			
Data source	Attribute quantity	Data type	Data structure
Basic claims data	4	Attribute field	Structured
MCGE report	3	Free-text	Semi-structured
Diabetes report	9	Free-text	Semi-structured
LIS	9	Attribute field	Structured

Table 6.1: Analysis of the data sources. Data sources used for the IE approach are printed in bold face.

Concentrating on the aggregation of *clinical data*, different data sources with relevant study attributes were analyzed and their underlying data structure explored. The results are summarized in Table 6.1, which shows that approximately 50% (12/25) of the attributes were documented within free text, and stored within a semi-structured data environment. These semi-structured data sources were investigated in this text mining study. Obviously, the appropriate extraction of attributes documented in clinical narratives is essential in an overall data aggregation system that handles clinical data which could be used in conjunction with biobanks. Further investigation showed that eight out of twelve free-text attributes were expressed as value/unit pairs, three got Boolean quantifiers ("Yes", "No") and one attribute got its value out of a standardized value set with three possible expressions ("steatosis hepatis", "no steatosis", "not possible")[1] (Table 6.2).

Based on patient codes and since this investigation is only interested in the attributes hidden within their clinical narratives, the MCGE and diabetes reports were exported from the CIS using a defined ETL workflow (Talend Open Studio for Data Integration [2]). It is important to mention that only those patients were selected for the training and test set where both report types (MCGE and diabetes) existed. This resulted in 78 patients for which both the MCGE and diabetes reports were available. Data from 39 patients were used as training data for the top-down development of a set of regular expressions per attribute. The remaining 39 patients formed the test set.

Gold standard

A gold standard for the IE approach was created by exporting the previously collected attributes from the electronic Study Documentation System (SDS) specifically designed for the cohort study. The SDS had been set up before the cohort study started so that

[1] An ultra sound examination can be used to diagnose a fatty liver (steatosis hepatis). Due to the fact that people undergoing this examination are obese (Body Mass Index ≥ 30) it is sometimes not possible to get valid results.

Attribute	Description	Value/Unit	Standardized Value Set	Boolean
height	Body height	•		
weight	Body weight	•		
BMI	Body Mass Index	•		
waist	Waist circumference	•		
hip	Hip circumference	•		
rrSystDiast	Systolic/Diastolic blood pressure	•		
dmII	Diabetes mellitus type II			•
familyHistory	Family history of obesity			•
hypertension	Hypertension			•
fibroScan	Fibroscan procedure	•		
iqr	Fibroscan inter quantile range	•		
us	Ultra sound procedure		•	

Table 6.2: Overview of the study-relevant attribute values embedded within semi-structured free text sources.

the physician responsible for the data acquisition could enter the relevant data collected from the patients into a structured form.

Evaluation metric

The performance of the IE system was assessed using the precision, recall and F-measure parameters described in Section 3.6.

Matching criterion. The gold standard was standardized with regard to the number of decimal places, however, this was not always the case in the text itself. For instance, a BMI of 40.2 was documented in the gold standard, but "40,196654" was found in the text. Here the extracted numerical string was rounded up or down to the number of decimal places as expected by the gold standard. Furthermore, the decimal separator (",", in German) was normalized.

Multiple expressions. Exactly one value was assigned to all attributes in the gold standard. The extraction routine produced an unordered set of one or more *unique* values. Two or more values were found when the same attribute occurred in the document more than once with different measurements, e.g. in the case of body mass or blood pressure.

6.3.2 Evaluation architecture

Systems. A CIS is usually composed of diverse subsystems, in which the structure of the data sources varies considerably, from highly structured data within a LIS, semi-structured data in clinical narratives, to unstructured data such as images stored in a picture archiving and communication system.

ETL. An ETL workflow exports the documents from the CIS where relevant study attributes were found to be embedded in the semi-structured text fields. Table 6.2

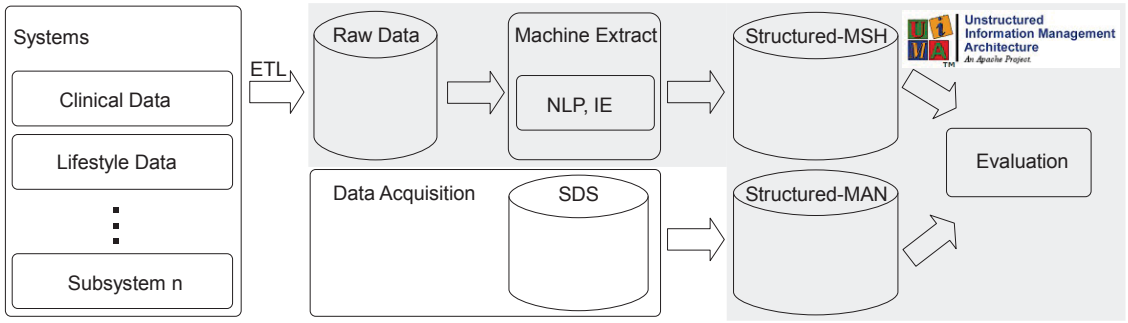


Figure 6.1: Evaluation architecture. Elements within the highlighted area are in the scope of Apache UIMA.

shows these attributes in the MCGE and diabetes reports. The documents were extracted based on a list of patient codes.

Raw Data. The result of the ETL workflow forms a pool of documents that contains relevant study attributes. MCGE reports are semi-structured documents containing XML tags. Diabetes reports are semi-structured PDF documents.

Machine Extract. Raw Data serves as input for the automated extraction for which Apache UIMA was used. The IE process was performed by using a set of regular expressions for each attribute, created from the training data. The rule-patterns were later applied to the test data set and performance values were estimated.

Structured-MSH. The result of the automatic extraction is a structured data set called Structured-MSH. Structured-MSH is a transient `jCAS` object within the UIMA architecture. It was used in the evaluation process.

Data Acquisition. Data Acquisition is the process of collecting and entering the relevant attributes of the patients into the SDS specified for this cohort. A data export of the SDS served as a gold standard (see Section 6.3.1) for the evaluation of the IE process.

Structured-MAN. Structured-MAN is an export of the SDS for this cohort in CSV-format, which was imported into a separate database for use in the evaluation pipeline.

Evaluation. The last step comprises the evaluation of the structured data from the IE process, in which the machine-extracted values are compared with the gold standard. The evaluation was done separately for the training and test data set.

6.3.3 Implementation aspects

UIMA is a well-established standard in industry and the scientific community for unstructured information processing (Section 2.4). In the following section some implementation

aspects are highlighted. The used parts of the UIMA framework (Figure 6.2) are described and combined with the evaluation architecture described in Figure 6.1.

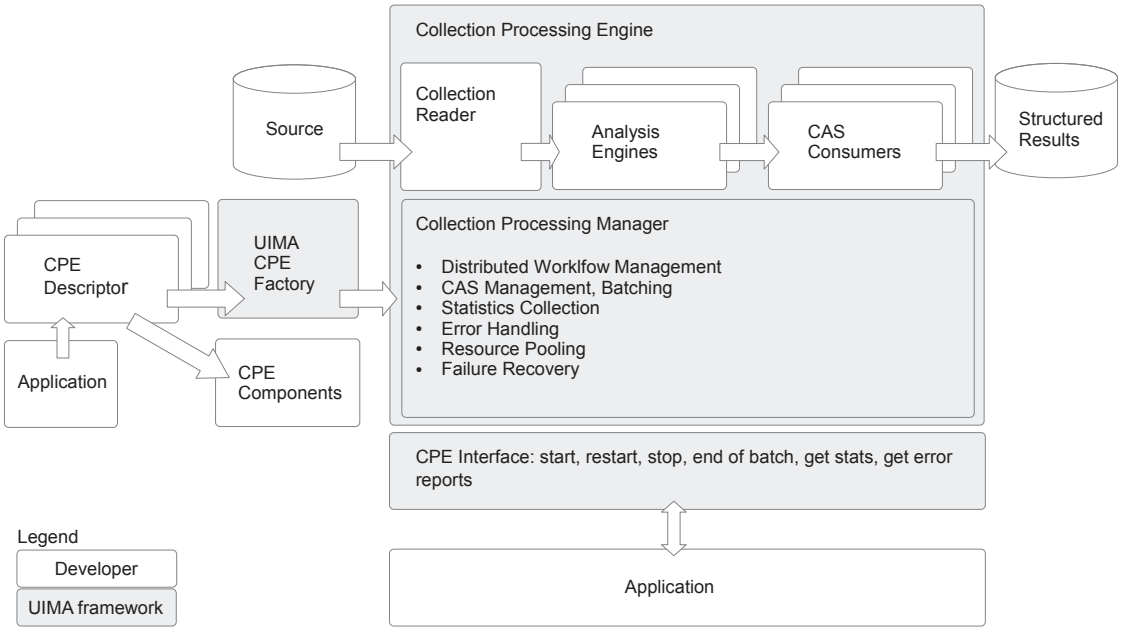


Figure 6.2: Apache UIMA [172].

Source. The document pool to be analyzed is Raw Data (Figure 6.1). It forms a pool of all diabetes and MCGE reports per patient in the study. A `CollectionReader` processes XML and pdf documents.

Collection Reader. Raw Data is read in by a `PDFCollectionReader` using iText [173]. This was because the diabetes reports are only available in PDF. Any existing diabetes and MCGE findings are summarized in a one document per patient basis which is subsequently analyzed in the pipeline.

Analysis Engine. For each attribute, a separate Analysis Engine was implemented. For some attributes text zoning was applied, in order to identify specific sections within the narrative, thus getting the correct attribute value pair after the processing chain. The Aggregated Analysis Engine (AAE), handling all attributes, maps to the Machine Extract component from Figure 6.1.

CAS Consumer. An implemented `EvaluationWriterCasConsumer` compares the structured, machine extracted attributes (Structured-MSH) to the manually collected (Structured-MAN) attributes, and evaluates the results of the entire extraction process. The reference gold standard was accessed from a MySQL database using the Hibernate [174] persistence framework.

Structured Results. Machine and manual collected results were not stored in a database but written pairwise into a text file for optional manual evaluation purposes, during

the last step of the evaluation architecture (Figure 6.1). The evaluation measures are calculated automatically at processing time by the IE engine.

6.4 Results and discussion

6.4.1 Regular expression analysis

Table 6.3 shows that text zoning was applied to a subset of attributes: such attributes generally appeared in the same paragraph within the narrative and were combined with a standardized text header; these headers, however, were occasionally modified or overwritten by users. By applying the regular expression `(?s)(Erhobene Befunde:.*?)(?=\n\s*\n[\wäöüÄÖÜß\s]*:\n)`, along with the information that the relevant section had a header of a given type, (e.g. "Erhobene Befunde:" (findings obtained)) the area for IE was narrowed down.

One regular expression `[Dd]iabetes [mM]ellitus [Tt]yp 2` was used for extracting the Boolean value for the existence of diabetes mellitus type 2; this was due to the fact that identical, highly standardized terms were always used, and therefore did not require the further consideration of linguistic variations. Another example was the extraction of body mass, for which the following three expressions were built together with text zoning due to the training set information: weightPattern: `Gewicht:\s*\d+[.,]?\d*\s*(kg)?`, valuePattern: `\d+[.,]?\d*`, unitPattern: `kg`. More effort for value extraction was required where linguistic variations of affirmed conditions had to be interpreted, e.g. for the ultrasound attribute `[eE]rhöhte Echodichte i.S.` (increased echo density in sono) or `ist vorhanden` (is present). Positive variations predominated, as a certain status in the investigated use case scenario was generally not documented if not existent.

The regular expressions strictly followed the information from the training data. No test set based information was used. Table 6.3 shows that the number of regular expressions applied, varied from only one (namely, that which did not use any additional section information from the narratives), up to 14 (those that exploited section annotations generated by the text zoning component).

6.4.2 Performance analysis

This section focuses on the performance of the IE system. It is highlighted whether errors occurred, where they originated from and subsequently categorized them. This allowed us to discover which attributes were difficult to extract and for which a minimum set of regular expressions were already sufficient for a good performance. The results are presented in Table 6.3.

An overall F-measure of 0.91 was achieved for the different attributes of the training set and an F-measure of 0.90 for the test set. A precision of less than one is generally the result of cases where the IE system retrieved several possible values ($max_{training} = 3, max_{test} = 4$) per attribute and patient. Out of 468 ($39 \cdot 12$) values from the gold standard, 433 cases

	Training			Test				
Attribute	Precision	Recall	F-measure	Precision	Recall	F-measure	Num RegEx	Text Zoning
height	0.89	0.95	0.91	0.94	1.00	0.96	3	Y
weight	0.83	0.90	0.86	0.86	0.95	0.88	3	Y
BMI	0.85	0.90	0.86	0.91	0.97	0.93	2	Y
waist	0.87	0.92	0.89	0.87	0.92	0.89	3	N
hip	0.87	0.92	0.89	0.89	0.92	0.90	3	N
rrSystDiast	0.91	0.95	0.92	0.90	0.95	0.92	3	N
dmII	0.97	0.97	0.97	0.97	0.97	0.97	1	N
familyHistory	0.95	0.95	0.95	0.82	0.82	0.82	4	Y
hypertension	0.67	0.69	0.68	0.72	0.74	0.73	3	N
fibroScan	0.97	0.97	0.97	0.90	0.90	0.90	8	N
iqr	1.00	1.00	1.00	1.00	1.00	1.00	8	N
us	0.97	0.97	0.97	0.87	0.87	0.87	14	Y
Mean	0.90	0.93	0.91	0.89	0.92	0.90		

Table 6.3: Information extraction evaluation results.

were found with the same values by the IE system using the training set, resulting in an overall recall of 0.93. The 35 mismatches fall into five distinct categories:

Disparate Values. The corresponding value of the attribute extracted by the IE system out of the clinical narratives differed from the value documented in the SDS. This could be due to an incorrect input into the clinical narratives or data acquisition for a single patient occurred at two different time points within both systems (CIS, SDS).

Machine IE. The automatic extraction approach revealed missing data within the SDS. Since it is a gold standard, it was assumed that missing data could not be collected from the patient during data acquisition for the cohort study.

Doc. Error. A typing error appeared in the routine documentation compared to the documentation in the SDS. A typing error can lead to a non-appropriate value for a certain attribute type expected to be in a certain range.

Interpr. Logic. The attribute variable and the corresponding value were misinterpreted due to a missing logic for value generation in the IE system.

Not Trained Pattern. The regular expressions were created in a way that they optimally retrieved information from the training set. In the test set new variations occurred for which patterns were not trained. Consequently, they remained unrecognized.

Table 6.4 shows that 49% of mismatches occurred due to differing information documented in the SDS and the clinical narrative. A high rate of evaluation errors were found because of attributes and their values, as extracted and interpreted from the IE system, revealed missing data within the SDS. The third important error category was due to missing interpretation logic for certain values. Table 6.4 shows that hypertension was responsible for all IE errors in this category. Hypertension was interpreted as existent if the attribute

Attribute	Disparate Values	Machine IE	Doc. Error	Interpr. Logic	Not Trained Pattern	Sum
height	2	0	0	0	0	2
weight	4	0	0	0	0	4
BMI	4	0	0	0	0	4
waist	3	0	0	0	0	3
hip	2	0	1	0	0	3
rrSystDiast	1	1	0	0	0	2
dmII	1	0	0	0	0	1
familyHistory	0	2	0	0	0	2
hypertension	0	8	0	4	0	12
fibroScan	0	1	0	0	0	1
iqr	0	0	0	0	0	0
us	0	1	0	0	0	1
Sum	17	13	1	4	0	35
Error Percentage	49%	37%	3%	11%	0%	100%

Table 6.4: Error analysis training data set.

value rrSystDiast was higher than 130. The rule actually applied by the physician was that hypertension was marked existent if the attribute value rrSystDiast was higher than 130, the patient took antihypertensive drugs, or there was arterial hypertension documented in the medical patient record. One typing error was found in the CIS regarding an attribute value. Furthermore, all relevant attributes and their values were recognized in the extracted clinical narratives in the training set; therefore there were no extraction errors due to a missing or overly strict regular expression. This is reasonable as the training data was used to set up the regular expression-based IE approach as accurately as possible.

Attribute	Disparate Values	Machine IE	Doc. Error	Interpr. Logic	Not Trained Pattern	Sum
height	0	0	0	0	0	0
weight	2	0	0	0	0	2
BMI	1	0	0	0	0	1
waist	3	0	0	0	0	3
hip	3	0	0	0	0	3
rrSystDiast	2	0	0	0	0	2
dmII	0	0	0	0	1	1
familyHistory	3	0	2	0	2	7
hypertension	0	8	0	2	0	10
fibroScan	2	2	0	0	0	4
iqr	0	0	0	0	0	0
us	0	2	0	0	3	5
Sum	16	12	2	2	6	38
Error Percentage	42%	32%	5%	5%	16%	100%

Table 6.5: Error analysis test data set.

The results for the test set are given in Table 6.3, with a recall of 0.92, estimating 430 out of 468 values documented within the SDS correctly, leaving 38 mismatches. Table 6.5 now shows a noticeable increase in the Not Trained Pattern error category, with us, familyHistory and dmII being the responsible attributes. The higher error rates were mainly due to linguistic variations that had to be interpreted for assigning a corresponding attribute value. Text patterns that had to be interpreted for the existence or absence of a

steatosis hepatis for the attribute us were, e.g., “vereinbar mit” (consistent with), “erhöhte Echodichte i.S. einer” (increased echogenicity in a sonogram), “Hinweise auf” (indication of), “geringgr.” (minor), “geringr.” (minor). More sophisticated NLP methods would have to be used rather than pure regular expression matching in order to improve correct value assignment [18, 40, 175–177]. In contrast, comparing Table 6.4 and Table 6.5 the absolute error numbers in the remaining error categories were approximately equal. A screenshot of the UIMA CAS Annotation Viewer GUI for manually checking the annotations made by the IE system is shown in Figure 6.3.

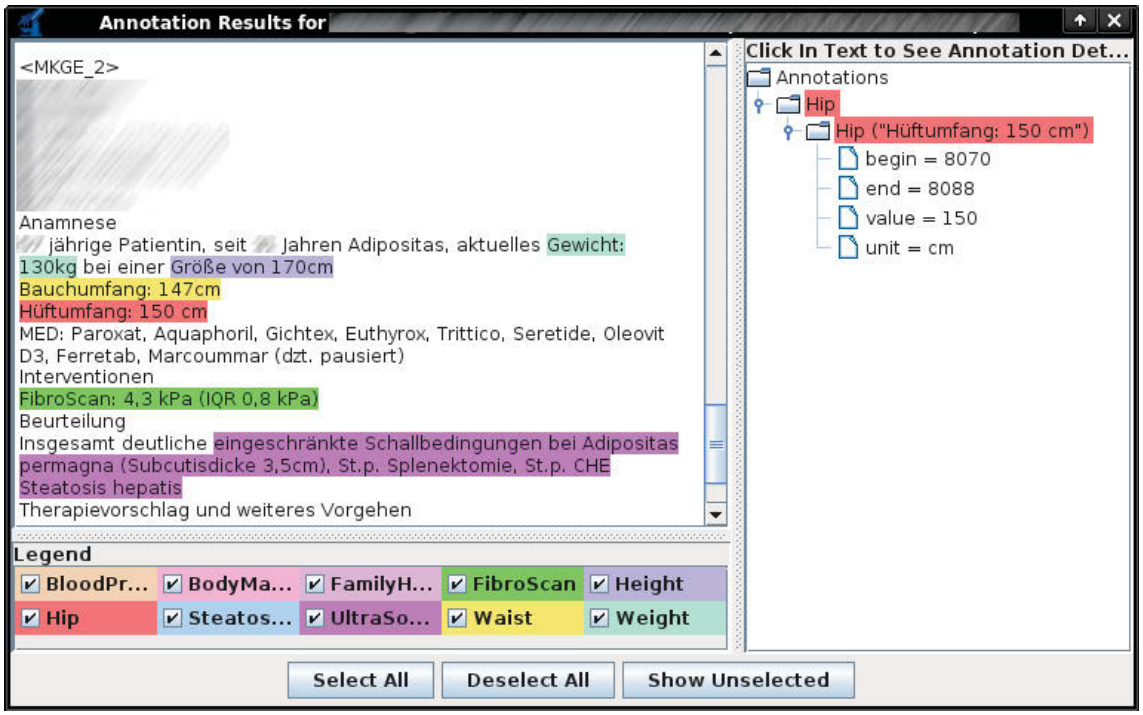


Figure 6.3: UIMA CAS Annotation Viewer GUI.

The following list gives a summary of the main pitfalls and challenges encountered, with an emphasis to the IE component:

- Typing errors.
  - “Hüftumfang: 13 cm” (hip circumference)
  - “Staeatosis hepatis” (steatosis hepatis, fatty liver)
- Inconsistency.
  - “Fibro-Scan”, “FIBROSCAN”, “Fibro Scan” (transient elastography, fibroscan)
  - “Grösse von 1.55m” versus “155 cm” (body height of)
- Redundancy.

  - Attributes were documented twice in different documents (with different values).

- Spelling variants for positive/negative patterns.
  - Pre-positive patterns: "geringgr." (minor), "ggr." (minor), "vereinbar mit" (compatible with)
  - Post-positive patterns: "ist vorhanden" (exists)
  - Pre-negative patterns: "kein Hinweis auf" (no evidence of)
  - Post-negative patterns: "ist nicht vorhanden" (is not present)

6.5 Conclusion and outlook

This work assessed the accuracy of the extraction of information from clinical narratives into a structured template. Based on a cohort study on metabolic syndrome the sources and the data structures in which relevant information was represented were analyzed. About 50% of the attributes relevant to the study were in semi-structured document templates. The Apache UIMA framework was used together with a rule based system using regular expressions as the core IE engine, specifically tailored to the content under scrutiny. An F-measure of 0.91 was obtained for the training set and an F-measure of 0.90 for the test set using the content in a SDS as gold standard. *Typing errors*, *inconsistency*, *redundancy* and *spelling variants* were identified as the main challenges for the IE approach proposed. In contrast, for variables contained in a quasi-standardized text format, a minimal set of regular expression was sufficient to obtain accurate extraction results.

There are increasing efforts in Europe [178] and the United States [179] to set up a collaborative biobank network. From an information management perspective CIS data quality is a major success factor, for the integration of clinical routine data and biobank sample data (disregarding the ethical issues that have to be considered). As well as the impact of data quality, three other aspects need to be stressed:

NLP. The results demonstrate that some attributes can be extracted with minimal effort. As a consequence, productive IE frameworks should be readily adjustable to specific information needs, accounting for special medical sub-language phenomena. For *typing errors* a text cleansing step prior to the NLP processing itself should be applied. *Inconsistency* in this case was handled by advanced regular expressions. A valid range set with respect to a given unit of measurement could further reduce false positives and value-unit pairs could be normalized to a pre-set standard. In this study, the *spelling variants* for a particular word meaning proved to be the most difficult to handle. They could be resolved by, e.g., either using handcrafted or openly available synonym lists or distributional semantics combined with edit distance measures.

Context. Context information would be an important information layer to address the problem of *redundancy* by applying data provenance information to an extracted

value. If the extraction strategy aims at finding all existing attributes and their values, such as in this use case, reliability information attached to the extracted value would be helpful, e.g. body height as reported by the patient vs. as measured by a nurse. Similarly, values from a quality controlled SDS would be more reliable than values extracted from text. This kind of quality annotation could support more differentiated retrieval scenarios, e.g. giving preference to either precision or recall.

Time. Important is the distinction between time stamps that refer to a patient centered event and time stamps that convey meta-information about a document, such as the creation date. Tools created for these tasks, such as the UIMA-based HeidelTime [180] annotator, are available and can be further adapted. However, this does not dispense with the need for carefully analyzing the way in which time references are handled in a class of documents and how they relate to certain patient-based events.

The investigations in this chapter have shown that it is important to find solutions for cohort building that lie between the documentation quality levels of *clinical documentation* and *documentation of clinical trials*, when it comes to distributed data aggregation and IE. Put simply, the higher the documentation quality with respect to the use of standardized documentation templates, the less sophisticated systems have to be built for extracting relevant information retrospectively. An example of a complex information need that would require a group of systems within a CIS to be accessed could be: "Do biobank samples exist that belong to smokers with a metabolic syndrome?" Regarding this investigation, three out of five values needed to diagnose a metabolic syndrome are in the LIS (triglycerides, HDL cholesterol, elevated fasting glucose). Two values are hidden in the free text of two different clinical documents (elevated waist circumference, elevated blood pressure). Finally, the list of patients with a metabolic syndrome would have to be merged with biobank and lifestyle data which was also discussed by Gostev et al. [181] in their implemented biobank sample management program SAIL. Furthermore completeness of information recorded in an EHR is an important aspect. The technical feasibility of the IE process does not mean that all relevant attributes needed for checking a study hypothesis are documented, as reported by Botsis et al. [161]. Cohort building as an example of the secondary use of clinical data is promising and increasingly requested. There is a trend to find out whether cohorts for retrospective or prospective studies can be reliably built based on routine documentation. Another reason is that clinicians are increasingly aware that routine documentation, often perceived as a burden, can produce new insights into patient groups. As recently stated by Hripcsak and Albers [182], unlocking information hidden in EHRs requires a compromise between a bottom-up and a top-down approach. Additionally a certain level of documentation quality, document structure, and use of standardized terminologies is needed. There is also a trade-off between, on the one hand, huge CIS systems that need to fulfill the legal requirements to store patient data for decades in diverse sub systems (which includes dealing with legacy data) and, on the other hand, the possibilities of novel frameworks and technologies for unstructured information processing. As a consequence, special purpose search servers implementing state-of-the-art technologies for selected CIS content could arise, bridging the gap between *patient-based storage systems* and *disease-related search systems*.

Chapter 7

Clinical Natural Language Processing

7.1 Introduction

For patients with numerous treatment episodes, the sum of their discharge summaries is an essential source of information about their current and past health problems and the progression of chronic diseases, encompassing signs, symptoms, allergies, diagnoses, medications, and procedures, embedded in contexts such as time or diagnostic certainty. Large quantities of documents and information can therefore accumulate in their EHR. To support a web-based patient-centered document search and navigation, a natural language processing toolset for document annotation (UIMA) and indexing (Solr), provided by the technology partner Averbis GmbH, was exploited and optimized for the text base under scrutiny [3]. One important part in clinical narrative processing is the correct handling of punctuations. In this context the focus was set on the period or full stop character (“.”).

This chapter is organized as follows: Section 7.1.1 gives a detailed problem analysis and description of period character disambiguation. Section 7.2 presents related work in the context of handling abbreviations within clinical narratives. The next Section 7.2 provides a detailed description of methods, language and data resources to find an adequate solution to the stated problem. Section 7.4 presents detailed insights into the results of handling the period character in clinical narratives in a proper way. The last Section 7.5 concludes the work presented in this chapter and gives an outlook to future work within this problem domain.

This chapter is based on Kreuzthaler and Schulz [183] following the copy right statement stated in Appendix B.2. The work received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement n^o 611388 - the SEMCARE project.

7.1.1 Problem analysis

The full stop or period character is known to be ambiguous. Besides its primary use as a sentence delimiter, it is often collocated with an abbreviation ("e.g.", "etc.", "Prof."). Periods also occur in numeric expressions ("13.2 mg") including dates ("24.1.2014"), as well as in a series of specialized names like file names ("readme.txt"), web addresses ("www.wikipedia.org"), or codes (e.g. International Classification of Diseases: "A01.9"). This is roughly true for all Western languages; yet minor variations exist between languages and dialects, e.g., the use of the period as decimal delimiter, its use in date and time formats, or the rules that guide its collocation with abbreviations.

A character-wise analysis of text allows for a distinction between period characters that are enclosed between two alphanumeric characters, and period characters that are followed by at least one, non-alphabetic character, such as a further punctuation sign, a space, tab or new line. The latter phenomenon is the focus of this study. Three cases are distinguished:

- Period characters that constitute the last character of an abbreviation.
- Period characters that follow non-abbreviated words and are sentence delimiters.
- Period characters that are part of an abbreviation and delimit a sentence.

In cases where the period is enclosed between two alphanumeric characters, it is considered an internal part of a token. This may be confounded with cases in which the space after a period is erroneously omitted, which masks token and sentence boundaries. However, the correction of such punctuation errors is considered as a separate problem outside the scope of this study.

If the roles of period characters are not appropriately clarified, sentences are split improperly, which has a severe impact on text analytics. In addition, if a system fails to identify abbreviations, their interpretation by mapping to full forms is impaired. Compared to fictional or journalistic texts, this distinction is particularly relevant for narrative clinical notes, in which periods are much more frequent [184]. Methods are investigated for identifying and classifying period characters in these text types, as a sub task of the so-called sentence boundary detection problem.

This clinical documentation use case focuses on text as it is typed into the computer by the physician at the point of care, or alternatively dictated and then processed by professional typists. In general, narratives constitute the most comprehensive and elaborate part of electronic patient records. Discharge summaries, in particular, constitute a rich abstract of those facts in a treatment episode that are considered relevant for decision making. Thus, discharge summaries are important vehicles for inter-physician communication, but they have also been increasingly valued as a rich source for the extraction of clinical information within so-called secondary use scenarios [185].

Clinical language is characterized, among other peculiarities like misspellings, punctuation errors and incomplete sentences (Section 2.2), by an abundance of acronyms and abbreviations [18]. This is why the focus is set on the use of the period character to distinguish between sentence limits and abbreviations. In German language abbreviations are nearly

mandatorily followed by a period such as “etc.”, in contrast to acronyms, which catch one’s eye by the collocation of several capital letters and, occasionally, digits, e.g., “EKG”. Non acronymic non-period abbreviations (like “Prof”) are generally disallowed. Physicians comply surprisingly well with this rule (in contrast to other editing errors they commit), and the exceptions are limited to a few frequent and short examples (e.g. “li” and not “li.” for “links” (left), “Supp” and not “Supp.” for “Suppositorien” (suppositories)).

The texts exhibited a tendency towards unusually lengthy abbreviations, chosen as a means to abbreviate long words (often single-word compounds) at the point where their (visual) completion by the reader can be taken for granted. Examples: “Penicillinallerg.” (“...ie”), “melanozyt.” (“...ische”), “paraffineingebett.” (“...et”). As long as such ad-hoc abbreviations are intuitively understandable, they are tolerated in medical notes, although they would never be admitted by medical publishers. Ad-hoc abbreviations are not lexicalized, but they commonly constitute substrings that are specific to a lexicon entry (albeit not necessarily to any determined inflection or derivation). In German, periods also frequently occur as markers for ordinal numbers, where similar ambiguities are observed. A snippet like “5.” may be read as a cardinal number followed by a sentence delimiter in “The first convulsion occurred at the age of 5.”. In “it was the 5. occurrence” it must be read as a ordinal number, at least in German, in which the period is a mandatory marker for ordinals, in default of special markers like 5^{th}, 5^{o} or $5^{ème}$. Finally, in “This fracture was his 5.”, the period has both roles. Into the concept of ordinals also certain date formats such as “3.5.” (in German, “the third of the fifth”) are included, in opposition to “3.5.2014” (in German, “the third of the fifth, two thousand and fourteen” - and not “fourteenth”). Due to the similarity to the phenomena of abbreviations, the concept of abbreviations to ordinal numbers is extended, arguing that “1.” is the abbreviation for “erst(e)(r)” (first), “2.” for “zweit(e)(r)” (second) and so on. The following example from a medical text exhibits numerous short forms, which will be analysed in more detail.

```
3. St.p. TE eines exulz. sek.knot.SSM (C43.5) li Lab. majus.
Level IV, 2,42 mm Tumordurchm.
```

In “3.” the period marks an ordinal number and also a sentence delimiter of the overall short sentence “Thirdly.”, introducing an enumerated list item; “St.p.” is the abbreviation of “Status post” (state after); “TE” is an acronym derived from “Totale Exzision” (total excision). “exulz.” like “Tumordurchm.” are ad-hoc abbreviations for “exulzerierendes” (fungating) and “Tumordurchmesser” (tumour diameter), respectively. “sek.knot.SSM” is an ill-formed agglutination of two abbreviations and one acronym. In correct writing, they should be separated by spaces (“sek. knot. SSM”). The abbreviation “sek.” (secondary) is a common, lexicalized one, whereas “knot.” (“knotig”, nodular) is, again an ad-hoc creation. “SSM” is an acronym for “Superfiziell Spreitendes Melanom” (superficial spreading melanoma). “C43.5” is a code from the International Classification of Diseases [186]. “Lab.” means “Labium”, a common anatomical abbreviation. “IV” is not an acronym, but a Roman number. “2,42” is a decimal number, which demonstrates that in German the period is not used as a decimal separator. Finally, the abbreviation “Tumordurchm.” demonstrates that the period plays a double role, *viz.* to mark an abbreviation and to conclude a sentence.

7.2 Related work

The detection of abbreviations and, which is closely related, the identification of the syntactic function of punctuation characters is important due to the frequency of both phenomena in clinical narratives [18].

There are several previous works on the disambiguation [187, 188] and normalization of short forms, with the goal to resolve the correct long form depending on its context. CLEF 2013 [189] started a task for acronym/abbreviation normalization, with a focus on mapping acronyms and abbreviations to concepts of the UMLS [190]. An F-measure of 0.89 was reported by Patrick et al. [191].

Xu et al. [184] tested four different abbreviation detection methods. The first one formed the baseline. Any unknown token within the narrative in comparison to a medical term list from MedLEE [192, 193] (containing 9,721 types plus Knuth's list of 110,573 American English words) was considered an abbreviation. The second was a rule-based approach that was customized by observing different admission notes, e.g. detecting whether the token contains a "." or "-" character. The third method used decision tree classifiers, using features of word formation and frequency, while the fourth method used additional features from other knowledge resources. Six admission noted formed the training set, four were used as a test set. The fourth method performed best with a precision of 0.91 and a recall of 0.80.

Due to the good performance measure of Xu et al. [184], Wu et al. [194] compared three machine learning methods (decision tree, SVM, random forest) for abbreviation detection. The training set comprised of 40 discharge summaries annotated by three experts; another 30 documents constituted the test set. Five different categories made up the full feature space: word formation, vowel combinations, related content from knowledge bases, word frequency in the overall corpus and local context. The random forest classifier performed best with an F-measure of 0.95 (precision 0.99, recall 0.91). A combination of classifiers lead to the highest F-measure of 0.96. In addition, Wu et al. [195] compared different clinical NLP systems on handling abbreviations in discharge summaries. MedLEE performed best with an F-score of 0.60 for all abbreviations. The implemented system, which addresses real-time constraints, is described in [196].

This investigation is only partially comparable, because it combines sentence delimitation with abbreviation detection. It is also peculiar due to the fact that the period character is mandatory as a non-acronym abbreviation marker in German, which causes severe disambiguation problems. In contrast to other work, this exploitation refrained from investigating acronyms. Maybe the notion of abbreviation is idiosyncratic, compared to the more general meaning of the term, especially regarding the English language, where abbreviations are often defined as any type of shortened term, including acronyms ("MI – Myocardial Infarction"), shortened words or phrases (e.g., "pt – patient"), and symbols (e.g., "eth – ethanol") [194]. Nevertheless, the distinction seems justifiable in the light of the particularities of German language, especially medical sub-language, for which – to the best of the author's knowledge – this investigation constitutes the first study on sentence delineation with the additional focus on abbreviation detection. In a preliminary study [197] the problem of sentence boundary detection together with abbreviation detec-

tion on similar texts had been addressed. An unsupervised statistical approach had been combined with a rule-based method for period character disambiguation. An accuracy of 0.93 for sentence boundary- and abbreviation detection had been obtained. Cases in which the periods were preceded by numerical characters had been excluded in that study, therefore the results are not fully comparable with the following investigation.

7.3 Materials and methods

7.3.1 Definitions and preprocessing

Based on a preliminary study [197], having applied a unsupervised statistical approach together with a rule-based method for the disambiguation of the period character within clinical narratives, the focus in this work is set on a supervised method exploiting SVMs for the two different tasks, *viz.* sentence delimitation and abbreviation detection. To this end, the formal notation introduced by Gillick [198] is extended together with that from Kiss and Strunk [199] to formalize the methodological approach on examples of the form "L• R", L• representing the left context token, • the period character ("."), and R the right context token. Note the token delimiter (here white space) between "L•" and "R". From this two tasks were derived:

1. Detection of abbreviations.
   P(a|"L• R")
2. Detection of sentence endings.
   P(s|"L• R")

A token is the result-output from a tokenizer. The straightforward Lucene [200] based WhiteSpaceTokenizer was applied. As a consequence, periods are always considered part of a token. All new line characters ("\n") are preserved before tokenization. As paragraph markers they will be used as features in the classification task. In addition, tokens containing only non-alphanumerical characters were merged with the preceding one. No manual cleansing was performed. Furthermore $norm(L\bullet) = L_{norm}\bullet$ is introduced as being a normalization by removing any non-word character except periods. Adjacent punctuations are merged. $norm(R) = R_{norm}$ replaces all non-word characters in R getting the *word* content. In the context of this work, (German) abbreviations are understood as being shortened words including a period character at the rightmost position, in contrast to acronyms which never include a period at their rightmost position.

7.3.2 Data

The data set was extracted using code-based search across all in- and outpatient discharge summaries from the dermatology department of the Graz University Hospital, covering the period between 01/2007 and 05/2013. The extraction was done using an ETL workflow with Talend Open Studio [2] and yielded 1,696 summaries. Both extraction and

anonymization were mandated by the data owner and conducted by the Scientific Service Area - Medical Data Management group, with the unique purpose to produce a non-identifiable medical corpus for advanced text mining studies. The anonymized patient summaries were divided into a training corpus (848 documents) and a test corpus (848 documents).

7.3.3 Gold standard

The sampling theorem with Chernoff bounds [201] was used to estimate a statistical representative sample size out of the training and test corpus with the following condition [202]:

$$n \geq \frac{2+\epsilon}{\epsilon^2} \ln \frac{2}{\delta} \tag{7.1}$$

With an accuracy of $\epsilon = 0.05$ and a confidence of $1 - \delta = 0.95$, $n = 3024$ text snippets were chosen as a representative gold standard size. The advantage of using the estimator theorem is its independence of the overall collection size for estimating a number of samples. By applying the estimator theorem it can be claimed that a feature estimate or representative syntactical pattern occurrence using the sampled corpus, with a probability of 95% is within $+\backslash-5\%$ of the truth. Therefore with this approach for sub sample size estimation it was attempted to fetch a significant amount of linguistic variations, which must be considered for interpreting the period character as an abbreviation or sentence delimiter (or both) and which allows generalizations from the experiments to the whole corpus.

By applying the theorem a reference standard was created through the random selection of 3024 text snippets for both the training and test set, centered on a period followed by a white space or newline, together with its left and right context (each 60 characters) from the sample texts. (For this experiment the sporadic cases are not considered in which spaces after periods were erroneously omitted). Two researchers rated each period character in the center of the snippet as functioning either as an abbreviation marker and/or sentence delimiter. As a measure of inter-rater agreement Cohen's kappa [15, 24] was calculated.

7.3.4 Language resources

Two German word lists were created and indexed: (i) an abbreviation-free medical domain dictionary (MDDict) with a high coverage of domain-specific words, excluding abbreviations, and (ii) a closed-class dictionary (CCDict) containing common, domain-independent word forms.

For MDDict, words were harvested from three sources: a free dictionary of contemporary German [203], a word list created out of raw text extracted from an old CD-ROM version of a medical dictionary [204], and medical texts and forum postings from a patient-centered website [205]. All tokens that ended with a period were discarded. The list comprised of about 1.45 million unique word types (the high number is due to inflectional/derivational variants and numerous single-word compounds), which were indexed with Lucene. Due

to possible punctuation errors (such as “etc” instead of “etc.”) it could not be guaranteed that the dictionary, at this step, was completely devoid of entries that would form a valid abbreviation if appended by a period. Therefore in a second step it was modified by two web resources containing German abbreviations [206, 207]. In total, about 5,800 acronym and abbreviation tokens were accumulated, of which terminal periods were stripped. Matching words were then removed from the initial dictionary.

For CCDict closed-class words are harvested from a German web resource [208], i.e. prepositions, determiners, conjunctions, and pronouns, together with auxiliary and modal verbs. The purpose of which was to have a comprehensive list of non-capitalized word forms, the capitalization of which always indicates the initial token of a sentence. The compilation of such a list benefits from a unique characteristic of the German language: namely, that all nouns are capitalized like proper names. Adjectives and full verbs may be capitalized, according to their syntactic role. Therefore, only German closed-class words follow capitalization patterns as in English, which warrants a high coverage for CCDict.

For the harvesting of the afore mentioned web resources Apache UIMA [39] was used, for which tailored CollectionReaders were implemented.

7.3.5 Support vector machines

The preference for SVMs (Section 3.1), is due to their known good performance on textual data [5] as well as their suitability for binary classification tasks. A linear kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ using LIBLINEAR [209] in combination with Weka [210] was exploited. In addition the data preprocessing methods described in [211] for generating the instances were applied, especially scaling the features to a range of [-1;1] and normalizing the feature vectors to unit length. The scaling of the test set was dimensioned according to the different value ranges in the training set. For selecting the optimal parameter C for the linear kernel a meta classifier (CVParameterSelection) was used varying the parameter C on a logarithmic scale [0.001;1000]. According to Joachims [212] “A small value for C will increase the number of training errors, while a large C will lead to a behavior similar to that of a hard-margin SVM”, therefore not allowing classification errors and having the tendency to over-fit.

Weka and its core methods for training and evaluation were encapsulated into a evaluation framework exploiting UIMA for rapid performance evaluation, getting fast access to true positives, false positives, true negatives and false negatives in the training and test set. Additionally, the functionality to obtain the top n most relevant features from the trained model, according to the method described by Guyon et al. [213] was implemented, which implies the use of a linear kernel. Taking this aspects into account, information relevance criteria can be formulated for each feature set combination. They are described in more detail in the next two sections.

7.3.6 Features for abbreviation detection

Statistical corpus features

Kiss and Strunk [199] used the log-likelihood ratio [214] for unsupervised abbreviation detection:

$$log\lambda = -2log(P(H_0)/P(H_A)) \tag{7.2}$$

H_0 is the hypothesis that the occurrence of a period is independent of the preceding word, H_A the alternative hypothesis *viz.* that it is not independent. The calculation of $log\lambda$ requires the corpus based frequency counts (Table 7.1) for every "L• R".

	L	$\neg L$
$\bullet$	$C(L_{norm}, \bullet)$	$C(\neg L_{norm}, \bullet)$
$\neg\bullet$	$C(L_{norm}, \neg\bullet)$	$C(\neg L_{norm}, \neg\bullet)$

Table 7.1: Corpus based frequency counts (C) required for $log\lambda$ calculation.

According to Dunning [215] one can "restate these counts slightly as the number of times the events occurred together" (in this notation Table 7.1 $C(L_{norm}, \bullet)$), "the number of times each has occurred without the other" (in this notation Table 7.1 $C(\neg L_{norm}, \bullet)$ and $C(L_{norm}, \neg\bullet)$) "and the number of times something has been observed that was neither of these events" (in this notation Table 7.1 $C(\neg L_{norm}, \neg\bullet)$). The implementation given in the Apache Mahout [216] package was exploited for calculating $log\lambda$ as well as building a frequency count map containing the different counts per $C(L_{norm}, \bullet)$ and $C(L_{norm}, \neg\bullet)$ collocation from which the other two frequency counts can be inferred. Finally, this feature set comprises $C(L_{norm}, \bullet)$, $C(\neg L_{norm}, \bullet)$, $C(L_{norm}, \neg\bullet)$, $C(\neg L_{norm}, \neg\bullet)$ and $log\lambda$.

Scaling combinations

As log-likelihood calculation tends to find all abbreviations but generally lacks precision [199], Kiss and Strunk applied different scaling factors (Equation 7.3 - 7.7) to $log\lambda$ for abbreviation [199] and sentence detection [217, 218] in combination with a threshold that had been defined by the authors after a series of experiments. In order to avoid setting a threshold arbitrarily, every possible scaling combination was generated of the factors described below and establishing each unique scaling combination as a separate feature. In combination with feature relevance ranking at the end, after training a model, the importance of scaling combinations for the classification performance can be assessed. The following single scaling functions form the base:

$$S_1(log\lambda, L_{norm}) : log\lambda \cdot e^{C(L_{norm},\bullet)/C(L_{norm},\neg\bullet)} \tag{7.3}$$

The scaling factor enhances the initial $log\lambda$ if the co-occurrence value $C(L_{norm}, \bullet)$ is greater than $C(L_{norm}, \neg\bullet)$.

$$S_2(log\lambda, L_{norm}) : log\lambda \cdot \frac{C(L_{norm}, \bullet) - C(L_{norm}, \neg\bullet)}{C(L_{norm}, \bullet) + C(L_{norm}, \neg\bullet)} \tag{7.4}$$

This scaling factor S_2 varies from -1 to 1 depending on co-occurrence counts of $C(L_{norm}, \bullet)$ and $C(L_{norm}, \neg\bullet)$. If $C(L_{norm}, \neg\bullet) > C(L_{norm}, \bullet)$ the scaling factor will be negative. If $C(L_{norm}, \neg\bullet) < C(L_{norm}, \bullet)$ the scaling factor will turn positive. The scaling factor equals zero if $C(L_{norm}, \neg\bullet) = C(L_{norm}, \bullet)$.

$$S_3(log\lambda, L_{norm}) : log\lambda \cdot \frac{1}{e^{wordLength(L_{norm})}} \tag{7.5}$$

This scaling factor punishes long words, based on the observation that most abbreviations are short.

$$S_4(log\lambda, L_{norm}) : log\lambda \cdot (\text{number of internal periods in } L_{norm} + 1) \tag{7.6}$$

This scaling factor gives an advantage to words that contain an internal period character, over words without internal period. The higher the number of internal periods in *word*, the higher is the chance that the word is an abbreviation.

$$S_5(log\lambda, L_{norm}) : log\lambda \cdot \frac{1}{wordLength(L_{norm})^{C(L_{norm}, \neg\bullet)}} \tag{7.7}$$

This scaling factor penalizes occurrences of L_{norm} exponentially which do not end with a period. This means that if they occur frequently, it is less likely to be an abbreviation also with respect to their length.

$$S_6(log\lambda, L_{norm}) : log\lambda + N(L_{norm}\bullet) \tag{7.8}$$

A sixth scaling function S_6 was introduced, which reflects the fact that most abbreviations are proper substrings of the shortened original word (e.g. "exulz." = "exulzerierend"), with $N(L_{norm}\bullet)$ being the sum of all found subword matches in $L_{norm}\bullet = (subword_1 \bullet subword_2 \bullet \ldots subword_n\bullet)$ for every $subword_i$ in a Lucene search result using MDDict. The reason why the last scaling function contains an addition, is to accommodate for cases where $C(L_{norm}, \bullet) < C(L_{norm}, \neg\bullet)$ even when L_{norm} is an abbreviation. These cases, for which the weighted $log\lambda$ is negative, could therefore be pushed to the positive side in the result of a strong S_6 attached at the end. The primary $log\lambda$ is modified by sequential composition of all possible variations of scaling factors (calculating the power set $P(S); S = \{S_1, S_2, S_3, S_4, S_5, S_6\}$), each resulting combination reflecting a feature.

Length features

Like Kiss and Strunk [218] the length of a word was considered as the count of all non-period characters, because internal periods should not have punishing effects.

$$wordLength(St.p.) = 3 \tag{7.9}$$

For building a descriptive abbreviation length statistic ($\mu = 5.8, \sigma = 4.4$) from the training corpus, those tokens were included that exhibited a significant $C(L_{norm}, \bullet)$ collocation ($p < 0.01$) and $C(L_{norm}, \bullet) > C(L_{norm}, \neg\bullet)$. Using this distribution the following length dependent features on L_{norm} were formulated:

Length of candidate. The absolute length of the abbreviation candidate, counting non-period characters.

Upper length border. The values of three right-tailed decision boundaries ($b_1 = \mu + 1.645\sigma$, $b_2 = \mu + 1.960\sigma$, $b_3 = \mu + 2.576\sigma$)

Binary decision rule. If the candidate is above one of the three different levels (b_1, b_2, b_3). Each decision result is a separate feature.

Mean minus length. The mean of the descriptive abbreviation length statistics minus the length of the candidate.

Word type features

L_{norm}, i.e. the word itself forms a feature. In order to keep the feature set small series of numerical digits were replaced with the character "d" within this feature set.

Rule-based features

Three different binary rules on L_{norm} were exploited:

Period character inside. The occurrence of at least one additional period character inside the candidate was assumed as an important information due to the fact internal periods are suggestive for abbreviations, especially when considering common date formats.

Contains numerical digit. This feature is assumed important, as containing digits have a relevant impact on whether the candidate should be classified as non-abbreviation or abbreviation. It is uncommon that tokens that contain digits are abbreviations, with the exception of ordinal numbers.

All upper case. Words that consist of upper case characters only are most likely acronyms or common words that were fully capitalized for some reason. The binary information of whether *all* characters within the candidate are upper case is used as a feature.

Dictionary-dependent features

This feature requires a dictionary lookup into MDDict, which is assumed to be devoid of abbreviations. If L_{norm} is found in the dictionary the feature value is set to 1, otherwise to 0.

7.3.7 Features for sentence detection

Abbreviation feature

Whether a candidate is an abbreviation or not constitutes the abbreviation feature. The best abbreviation classification SVM model and its feature combinations was taken and applied to "L• R". If classified as an abbreviation it should favor the decision against sentence delimitation, as most abbreviations tend to appear within and not at the end of a sentence.

Length features

The same length features as described in Section 7.3.6 were applied because length information of the right context R_{norm} was considered to have a relevant impact on sentence delineation.

Rule-based features

Four different binary rules were exploited. The first three were intended to introduce right-context-based abbreviation information to the sentence delimiter decision. The last rule is a direct sentence delimiter rule.

Period character inside. This is assumed to be an important information due to the fact that most tokens containing an internal period are abbreviations (Applied to R).

Contains numerical digit. The feature is assumed important, as containing digits have a relevant impact whether the candidate should be classified as a non-abbreviation or abbreviation (Applied to R_{norm}).

All upper case. Capitalized words are acronyms or emphasized words. If followed by a period they should be classified as non-abbreviations (Applied to R_{norm}).

Capitalization. The capitalization of the first character of R_{norm} is a good indicator for sentence delimiters, because new sentences generally begin with an upper case character.

Word type features

This feature set was generated in the same way as described in Section 7.3.6, i.e. using word type information. High performance values for sentence delineation with this feature set alone was reported by Gillick [198] therefore it was included as one feature type in this enhanced SVM approach on clinical narratives.

Right context word type features

This feature set was generated in the same way as described in the previous section, the only difference being that here it was applied to the right context R_{norm}. An n-gram representation was not done because this still allows the exploitation of some right context type information R_{norm}, in the case that the type information of L_{norm} is missing, or vice versa. The possibility of using R_{norm} and L_{norm} as a combined feature set has the advantage that it is not as strict as a stand-alone bi-gram (L_{norm}, R_{norm}) feature set. However, the bi-gram information is encoded if it exists, and the feature set is kept small.

Text formatting features

In well-formatted text a new line character after a period marks the end of a paragraph. Therefore, the period here can generally be assumed to play the role of a sentence delimiter, because sentences never span across paragraphs. This investigation could not strictly apply this rule, as parts of the clinical narratives under scrutiny were fragmented with new line characters "\n", as a side effect of the process that imported the narratives from the CIS, a phenomenon which is also well known when extracting raw text from PDF sources. As a consequence, only double new line characters could be safely considered as paragraph markers. Nevertheless the impact of this feature on the sentence detection task is investigated. The following features were formulated:

Single new line. The feature is set true if R starts with a single new line.

Double new line. The feature is set true if R starts with a double new line.

No new line. The feature is set true if R starts with no new line.

Language-dependent features

Similarly to Section 7.3.6 a lexicon lookup of R_{norm} in MDDict was performed in order to decide whether R_{norm} existed in the harvested dictionary. It is hypothesized that this is also important for sentence detection because it seems that a sequence of two abbreviations normally occurs within the same sentence.

7.4 Results

The evaluation results are presented in this section, starting with 10-fold cross validation on the training set. Afterwards the trained model was used for the performance evaluation of the test set. Results are provided as unweighted micro-averaged F-measures, as recommended by Manning et al. [11].

A Cohen's kappa of 0.98 of the gold standard annotations clearly reflects the fact that both abbreviation and sentence detection are easy tasks for human raters. By identifying the top 10 relevant features depending on the formed feature sets from the trained model, the impact of these features on the classification tasks could be stated. Significance tests (chi-squared test, $p < 0.05$) were applied with respect to the baseline and on different ranked feature set combinations.

7.4.1 Results of abbreviation detection

As a baseline, a straightforward decision algorithm was chosen: if the abbreviation candidate is followed by a lower case character it is classified as abbreviation, otherwise as non-abbreviation.

Method	BL	1	2	3	4	5	6
micro-avg. $F_{1\,Training}$	0.62	0.60*'	0.73*'	0.83*'	0.83*	0.83*	0.93*'
micro-avg. $F_{1\,Test}$	0.60	0.60	0.70*'	0.81*'	0.83*	0.84*'	0.92*'

Table 7.2: Abbreviation detection. Evaluation performance per feature set (1 Rule-based features; 2 Statistical features; 3 Scaling features; 4 Language-dependent features; 5 Length features; 6 Word type features). * significant difference to baseline (BL) ($p < 0.05$), ' significant difference to predecessor ($p < 0.05$)

First the feature sets for abbreviation detection were evaluated in isolation, of which the achieved performance values are depicted in Table 7.2. The following exemplification refers to the training set. The rule-based features showed poorer performance, also significantly inferior to the very straightforward baseline. Nevertheless, within this feature set (Table 7.3), the feature *Contains period* appears as the most relevant feature for this task. Following the rule-based features, the statistical feature set has, in isolation, a micro-averaged F-measure of 0.73 with the simple frequency count $C(L_{norm}, \bullet)$, listed as the most relevant within this feature set. After the rule-based feature set, the scaling, language-dependent and length feature sets achieve, in isolation, the same performance of 0.83. In this setting, interestingly, calculation-intensive statistical feature sets (scaling feature set, length feature set) result in roughly the same performance as a simple dictionary lookup. A respectable performance is achieved using only word type features (L_{norm}) yielding an F-measure of 0.93. The top 10 features within this set reflect the most common abbreviations within the corpus under scrutiny (Table 7.3). Only slightly lower performance values were achieved for the test set.

Top 10	1	w^2	2	w^2	3	w^2
1	Contains period	0.30	$C(L_{norm}, \bullet)$	1.34	S_2	3897.48
2	All upper case	0.02	$log\lambda$	0.80	S_3	3222.35
3	Contains digit	0.01	$C(L_{norm}, \neg\bullet)$	0.43	S_4	2592.76
4	-	-	$C(\neg L_{norm}, \neg\bullet)$	0.31	S_2, S_3	2329.77
5	-	-	$C(\neg L_{norm}, \bullet)$	0.19	S_4, S_5	847.88
6	-	-	-	-	S_5	706.98
7	-	-	-	-	S_2, S_4, S_5	511.38
8	-	-	-	-	S_2, S_5	412.86
9	-	-	-	-	S_3, S_4	204.80
10	-	-	-	-	S_2, S_3, S_4	139.36
Top 10	4	w^2	5	w^2	6	w^2
1	$\in$ MDDict	0.34	LT border b_2	16.15	St.p.	409.58
2	-	-	LT border b_1	16.15	Amb.	409.51
3	-	-	LT border b_3	16.15	o.B.	409.09
4	-	-	LT	8.74	re.	407.87
5	-	-	Mean-LT	8.74	Z.n.	407.35
6	-	-	$> b_1$	0.54	li.	407.28
7	-	-	$> b_3$	0.16	ca.	407.00
8	-	-	$> b_2$	0.10	unauff.	406.94
9	-	-	-	-	bds.	406.19
10	-	-	-	-	Pat.	405.75

Table 7.3: Abbreviation detection. Top 10 feature rankings per feature set (1 Rule-based features; 2 Statistical features; 3 Scaling features; 4 Language-dependent features; 5 Length features; 6 Word type features). Length (LT); w^2: Weight based feature relevance criterion.

Method	BL	[1]	[1-2]	[1-3]	[1-4]	[1-5]	[1-6]
micro-avg. $F_{1\,Training}$	0.62	0.60*’	0.71*’	0.86*’	0.88*’	0.95*’	0.97*’
micro-avg. $F_{1\,Test}$	0.60	0.60	0.71*’	0.83*’	0.86*	0.93*	0.95*’

Table 7.4: Abbreviation detection. Evaluation performance combining feature sets stepwise according to their standalone performance (1 Rule-based features; 2 Statistical features; 3 Scaling features; 4 Language-dependent features; 5 Length features; 6 Word type features). * significant difference to baseline (BL) ($p < 0.05$), ’ significant difference to predecessor ($p < 0.05$)

After the evaluation of each feature set in isolation these sets were combined stepwise, evaluating their performance (Table 7.4). Starting with the rule-based feature set and adding the statistical feature set, achieved an F-measure of 0.71. By combining these two feature sets a lower performance was obtained than using only the statistical set in isolation, comparing Table 7.2 with Table 7.4 for the training set. In the next step the scaling features were added, yielding in combination, an F-measure of 0.86 which is higher compared to the scaling features in isolation. Interestingly, when analysing this combined set (Table 7.5), the top 10 features were constituted by only scaling combinations and $log\lambda$. The same was true when adding the language-dependent set, with only the ranking being different. Nevertheless when adding the language-dependent feature set a performance gain up to an F-measure of 0.88 was achieved for the training set. After introducing the length features set, an F-measure of 0.95 was obtained, and, finally together with the word type features, the highest performance of 0.97 for the training set (Table 7.4) was achieved. The final top 10 features of this set are shown in Table 7.5. It is remarkable that within this ranking *no language-dependent features exist*, but at least one feature belonging to the other feature sets. An F-measure of 0.95 is achieved on the test set by combining all features.

Top 10	1	w^2	[1-2]	w^2	[1-3]	w^2
1	Contains period	0.30	Contains period	0.35	S_2	5885.83
2	All upper case	0.02	$C(L_{norm}, \bullet)$	0.18	S_3	4855.66
3	Contains digit	0.01	$log\lambda$	0.13	S_4	1999.51
4	-	-	$C(\neg L_{norm}, \neg\bullet)$	0.12	S_2, S_3	1798.60
5	-	-	$C(\neg L_{norm}, \bullet)$	0.09	$log\lambda$	1180.39
6	-	-	$C(L_{norm}, \neg\bullet)$	0.09	S_5	894.98
7	-	-	All upper case	0.02	S_4, S_5	715.70
8	-	-	Contains digit	8.16E-5	S_2, S_5	617.98
9	-	-	-	-	S_2, S_4, S_5	474.86
10	-	-	-	-	S_3, S_4, S_5	256.81
Top 10	[1-4]	w^2	[1-5]	w^2	[1-6]	w^2
1	S_2	1063.78	S_5	1027.15	LT	952.62
2	S_3	962.33	S_4, S_5	914.02	Mean-LT	952.62
3	S_2, S_3	507.82	S_2, S_5	610.69	All upper case	549.64
4	S_4	391.68	S_2, S_4, S_5	527.28	S_3, S_4, S_5	529.85
5	S_3, S_4, S_5	379.70	S_2	463.94	S_3, S_5	521.60
6	S_3, S_5	325.68	S_3, S_4, S_5	274.81	erforderl.	403.54
7	S_5	265.62	S_3, S_5	253.30	pathol.	392.23
8	S_4, S_5	222.55	Mean-LT	145.91	verschiebl.	375.40
9	$log\lambda$	143.67	LT	145.91	d-lat.	358.11
10	S_2, S_5	129.90	S_2, S_4	90.13	entzündl.	345.21

Table 7.5: Abbreviation detection. Top 10 feature rankings per feature set (1 Rule-based features; 2 Statistical features; 3 Scaling features; 4 Language-dependent features; 5 Length features; 6 Word type features). Length (LT); w^2: Weight based feature relevance criterion.

7.4.2 Results of sentence detection

The baseline algorithm for sentence detection analyses the capitalization status of R_{norm}. Only if capitalized, "L• R" is classified as sentence delimiter.

Method	BL	1	2	3	4	5	6	7
micro-avg. $F_{1\ Training}$	0.78	0.58*'	0.76*'	0.79*'	0.79*	0.82*'	0.90*'	0.92*'
micro-avg. $F_{1\ Test}$	0.75	0.60*'	0.74*'	0.78*'	0.81*'	0.77*'	0.87*'	0.92*'

Table 7.6: Sentence detection. Evaluation performance per feature set (1 Language features; 2 Rule-based features; 3 Text format features; 4 Word length features; 5 Right context word type features; 6 Word type features; 7 Abbreviation feature). * significant difference to baseline (BL) ($p < 0.05$), ' significant difference to predecessor ($p < 0.05$)

As depicted in Table 7.6, the language features and the rule-based features alone performed significantly worse than the baseline. Interestingly, using the text format features, even though the texts under scrutiny were heavily contaminated by new line characters, a performance above the baseline was obtained for the first time. The feature relevance ranking at this stage is shown in Table 7.7. There was no significant performance difference using the word length features in isolation compared to the text formatting features for the training set. The right context word type feature set based on R_{norm} performed worse than the word type based features using L_{norm}. The most important features in this set are shown in Table 7.8. The best performance was observed for the standalone feature set evaluation using only the information specifying whether "L• R" is an abbreviation or not, using the optimized SVM for abbreviation detection resulting in an F-measure of 0.92 for sentence delineation. The same performance was obtained on the test set. This reflects the important influence of abbreviation detection on sentence delineation.

Top 10	1	w^2	2	w^2	3	w^2
1	$\in$ CCDict	0.07	Capitalization	1.84	No "\n"	0.32
2	$\in$ MDDict	2.15E-3	All upper case	0.54	Double "\n"	0.06
3	-	-	Contains digit	0.27	Single "\n"	0.03
4	-	-	Contains period	1.59E-5	-	-
5-10	-	-	-	-	-	-

Table 7.7: Sentence detection. Top 10 feature rankings per feature set (1 Language features; 2 Rule-based features; 3 Text format features). w^2: Weight based feature relevance criterion.

As in Section 7.4.1 a stepwise combination of feature sets was performed in order to gain insight into their combined performance. The first positive significant impact on classification performance in comparison to the stand alone evaluation was achieved, when combining the first three feature sets, reaching an F-measure of 0.88 for the training set (Table 7.9). The feature relevance ranking at this point is depicted in Table 7.10.

Top 10	4	w^2	5	w^2	6	w^2	7	w^2
1	LT border b_2	60.82	Die	121.98	St.p.	415.20	Abbr	0.54
2	LT border b_1	60.82	für	121.03	Amb.	410.82	-	-
3	LT border b_3	60.82	TE	94.84	ca.	402.62	-	-
4	LT	1.98	Keine	94.09	Pat.	401.16	-	-
5	Mean-LT	1.98	Sono	83.67	max.	397.93	-	-
6	$> b_2$	0.13	Der	80.47	Z.n.	392.47	-	-
7	$> b_3$	0.04	CT	77.13	st.p.	390.62	-	-
8	$> b_1$	2.27E-4	E-Nr	75.40	n.	378.70	-	-
9	-	-	Im	71.92	St.	377.24	-	-
10	-	-	Am	66.45	bzw.	368.27	-	-

Table 7.8: Sentence detection. Top 10 feature rankings per feature set (4 Word length features; 5 Right context word type features; 6 Word type features; 7 Abbreviation feature). Length (LT); w^2: Weight based feature relevance criterion.

Adding stepwise word length features, right context word type features (R_{norm}), word type features (L_{norm}), and finally the abbreviation information, an unweighted micro-averaged F-measure of 0.97 was obtained for sentence detection using the training set and an F-measure of 0.94 for the test set (Table 7.9). Table 7.11 documents interesting insights into the most relevant features of the top performing model. The information whether "L• R" is an abbreviation or not is the most important one, followed by upper case information relating to R_{norm}. The remaining top 10 features are a combination of word type features (right and left context) and text format features. It is plausible that the text format features convey important information, but it has to be emphasized that the single occurrence of a new line is not in the top features anymore. Due to the contamination of the text sample in use by line breaks, this feature does no longer reliably predict sentence boundaries. This information has been automatically induced from the parameter and the feature-optimized SVM. The remaining important text format features for sentence detection are Double "\n", marking and end of a text passage, and No "\n".

Method	BL	[1]	[1-2]	[1-3]	[1-4]	[1-5]	[1-6]	[1-7]
micro-avg. $F_{1\,Training}$	0.78	0.58*'	0.76*'	0.88*'	0.92*'	0.95*'	0.96*'	0.97*'
micro-avg. $F_{1\,Test}$	0.75	0.60*'	0.75'	0.86*'	0.91*'	0.93*'	0.94*'	0.94*

Table 7.9: Sentence detection. Evaluation performance combining feature sets stepwise according to their stand alone performance (1 Language features; 2 Rule-based features; 3 Text format features; 4 Word length features; 5 Right context word type features; 6 Word type features; 7 Abbreviation feature). * significant difference to baseline (BL) ($p < 0.05$), ' significant difference to predecessor ($p < 0.05$)

Top 10	[1]	w^2	[1-2]	w^2	[1-3]	w^2
1	$\in$ CCDict	0.07	Capitalization	2.67	Capitalization	1.54
2	$\in$ MDDict	2.15E-3	All upper case	0.47	No "\n"	1.09
3	-	-	$\in$ CCDict	0.43	$\in$ CCDict	0.58
4	-	-	Contains digit	0.21	Double "\n"	0.48
5	-	-	Contains period	0.02	All upper case	0.17
6	-	-	$\in$ MDDict	8.32E-4	Single "\n"	0.11
7	-	-	-	-	Contains digit	0.07
8	-	-	-	-	$\in$ MDDict	0.03
9	-	-	-	-	Contains period	0.01
10	-	-	-	-	-	-

Table 7.10: Sentence detection. Top 10 feature rankings per feature set (1 Language features; 2 Rule-based features; 3 Text format features). w^2: Weight based feature relevance criterion.

Top 10	[1-4]	w^2	[1-5]	w^2
1	Capitalization	11.34	LT	674.21
2	LT	10.52	Mean-LT	674.21
3	Mean-LT	10.52	Capitalization	637.85
4	No "\n"	4.82	Rippenanteile$_{RC}$	627.54
5	$\in$ CCDict	4.08	Lymphknoten$_{RC}$	356.25
6	Double "\n"	3.77	Double "\n"	336.64
7	All upper case	0.97	Lungengerüstzeichnung$_{RC}$	332.50
8	Contains digit	0.71	Integument$_{RC}$	321.86
9	$> b_3$	0.31	No "\n"	300.18
10	Contains period	0.19	Normale$_{RC}$	277.68
Top 10	[1-6]	w^2	[1-7]	w^2
1	Capitalization	971.25	Abbreviation	1326.41
2	Mean-LT	840.45	Capitalization	867.06
3	LT	840.45	o.B.	382.83
4	Double "\n"	341.46	Double "\n"	374.57
5	No "\n"	324.25	No "\n"	364.32
6	o.B.	259.13	bds.	282.13
7	Rippenanteile$_{RC}$	254.91	CT$_{RC}$	266.54
8	mitresez.	254.91	Leberlappen$_{RC}$	225.08
9	CT$_{RC}$	251.41	A.	206.77
10	Leberlappen$_{RC}$	236.26	Narbige$_{RC}$	191.01

Table 7.11: Sentence detection. Top 10 feature rankings per feature set (1 Language features; 2 Rule-based features; 3 Text format features; 4 Word length features; 5 Right context word type features (RC); 6 Word type features; 7 Abbreviation feature). Length (LT); w^2: Weight based feature relevance criterion.

7.5 Conclusion and outlook

This investigation presented and evaluated a supervised machine learning approach using a SVM exploiting a linear kernel for abbreviation detection and sentence delineation in German-language medical narratives. The UIMA framework was used in conjunction with Weka. A modular evaluation framework was created in order to gain insight in different classification settings and feature relevance rankings. Exploiting this framework, for abbreviation detection an unweighted micro-averaged F-measure of 0.97 for the training set and an F-measure of 0.95 for test set based evaluation was achieved. For sentence boundary detection an unweighted micro-averaged F-measure of 0.97 for training set based evaluation and an F-measure of 0.94 using the test set can be reported. This is a comparable performance to the maximum entropy based sentence detection tool implemented within OpenNLP [74], exploited by cTakes [40] (sentence boundary detector accuracy=0.95). Both results are remarkable as clinical narratives have specific idiosyncrasies (Section 2.2), and are thus quite distinct from the proof-read content of textbooks and scientific papers [219].

Future work may explore how the achieved performance for abbreviation detection and sentence delineation can be enhanced by exploring additional feature sets, in order to minimize false positives and false negatives. This could be done by exploiting n-gram information, expanded context information, by additional corpus-based statistical features, or by word formation features as described by Wu et al. [194]. Due to the fact that comparable results to the cTakes sentence detection tool (applied to English clinical text) using OpenNLP were achieved, a direct comparison between the approach presented here, and a re-trained version of the OpenNLP sentence detection tool for German texts used in this investigation would be interesting for a supervised approach in general. Additionally, an enhanced version of the preliminary approach described in [197] could be further evaluated. Furthermore, the applicability to other clinical subdomains would be of interest, as different document types (e.g. dermatology clinic notes, neurology clinic notes) form distinct sublanguages, according to Friedman [220] and Patterson et al. [4]. Interinstitutional and interregional evaluations (e.g. comparing Austria, Germany, Switzerland for the German speaking community) could be investigated, in order to obtain more generally applicable NLP pipelines for medical document processing and to identify the needs for customization. Further work may also propose additional features that are language-independent and do not rely on language-specific dictionaries or rules. Language-independent implementations, also considering real-time constraints in a clinical setting, could further improve current clinical NLP frameworks, such as cTakes [40] or MedKAT [221] for the non-English clinical NLP community.

Chapter 8

Conclusion and Outlook

In this thesis different scenarios for language technologies in a clinical environment were investigated. The focus was placed on the processing of semi-structured clinical narratives, addressing several peculiarities of clinical language. Motivated by clinical IR use cases, language technologies were applied to four different scenarios.

The first scenario targeted the classification of clinical documents. The use case was motivated by the fact to give a doctor a hint whether a report is about an inflammation or neoplasm, within long list based views. In this scenario, morphosemantic normalization (using a subword thesaurus) on the one hand and character n-gram decomposition on the other hand were investigated with regard to classification performance of a parameter optimized support vector machine exploiting a linear kernel. Additionally, different weighting schemes were applied to the feature vectors (binary, term frequency, term frequency-inverse document frequency, I(n)B2). A micro-averaged F-measure of at least 0.95 could be reached for both classification tasks using just a simple binary weighting scheme together with 4-gram models. The use of different weighting schemes had less impact than expected. Without n-gram modeling the use of a morphosemantic normalization component improved classification performance but was dispensable if n-grams were applied. Calculation of expensive weighting schemes (I(n)B2) and pre-processing methods (morphosemantic normalization) had only a very small performance gain compared to simple straightforward tokenization, 4-gram modeling and the application of a binary weighting scheme ($\Delta_{F_1} = 0.002$ for inflammation detection and $\Delta_{F_1} = 0.012$ for neoplasm classification). The supervised machine learning method could not reach the classification performance of a rule-based knowledge engineering approach exploiting regular expressions, for which rules have iteratively been handcrafted and which reached an impressive F-measure of 0.98 [61, 62, 101], but the achieved performance is noticeable.

The second scenario was about clinical IR. Corresponding to typical information needs of physicians, a TREC style based IR evaluation scenario was established based on 26 information needs together with a gold standard. The main research question was to test IR models exploiting distributional semantics. LSA was combined with morphosemantic analysis, applied to three different tokenized versions of the original texts: first the Apache Lucene SnowballAnalyzer for German, second, tokens gained from morphosemantic analysis and third a combination of both. The binary, term frequency, term frequency-inverse

document frequency and I(n)B2 weighting schemes were applied to the term-document matrix. Each of the 26 information needs was transformed into a keyword-based search, according to three language registers, simulating the medical background knowledge of a user: language register *Laypeson*, language register *Netdoktor*, and language register *Expert*. According to the language register in use different keywords representing the information need were used. MAP and Precision at k were the evaluation metrics for ranked retrieval results. Aggressive dimension reduction to a size of 15, together with the I(n)B2 weighting scheme and the use of the Netdoktor language register yielded the best performance of a MAP of 0.55. A high degree of dimension reduction in general had a positive impact on retrieval performance as well as the different weighting schemes applied. In accordance to literature, I(n)B2 for this clinical IR scenario had the highest relevant impact. Queries using expert language achieved the highest performance values if no morphosemantic component was applied. In contrast, non-expert language queries benefitted from morphosemantic pre-processing, achieving retrieval results similar to expert language. The computational costs of this statistical approach are very high (processing the SVD) and the difficult estimation of the optimal degree of dimension reduction makes this statistical retrieval approach alone less applicable in a clinical IR scenario.

The third scenario investigated IE from clinical narratives into a structured template in order to assess whether clinical texts constitute appropriate sources from which structured information for clinical research can be reliably extracted. A cohort study on the metabolic syndrome was the use case for which the data sources- and structures that contain relevant information were analyzed. 50% of the attributes relevant to the study were found in semi-structured document templates. The Apache UIMA framework was used together with a top-down rule-based approach using regular expressions as the core IE engine, specifically tailored to the content under scrutiny. The prototypical IE system achieved an F-measure of 0.91 (precision = 0.90, recall = 0.93) for the training set and an F-measure of 0.90 (precision = 0.89, recall = 0.92) for the test set using the content in a SDS as gold standard. *Typing errors*, *inconsistency*, *redundancy* and *spelling variants* were identified as the main challenges for the proposed IE approach. In contrast, for variables contained in a quasi-standardized text format, a minimal set of regular expression was sufficient to obtain accurate extraction results. The results and the literature review both suggest that a regular expression based knowledge engineering approach optimized on the use case can be used to extract structured relevant study attributes from clinical narratives. Together with UIMA AS [222] the process could be parallelized for millions of clinical documents.

The fourth scenario uses NLP to address patient-based document search. This use case is of particular interest for patients with chronic diseases and numerous treatment episodes, for which large quantities of electronic documents accumulate in intra- and extra-institutional health record systems (with ELGA being a future application of the latter). A web based semantic clinical report navigator prototype was implemented [3] to support quick navigation and search within the whole document corpus related to one patient. Semantic relations and lexical variants should be taken into account, as well as a recall oriented customization of an NLP pipeline had to be made. One important functionality, processing clinical narratives is the detection of sentences boundaries and the identification of abbreviations. In German language, the period character plays a double role as sentence delimiter, abbreviation marker, or both. For this reason two support vector machines

exploiting a linear kernel where trained (3024 annotated text snippets) in combination with a rich feature engineering task. Test-set (3024 annotated text snippets) based micro-averaged F-measure evaluation showed a performance 0.95 for abbreviation detection and 0.94 for period character centered sentence boundary detection. Feature relevance ranking showed that the machine learning approach was able to automatically infer human recognizable relevant features e.g. for sentence boundary detection. In this case one of the most important features automatically inferred were if the period character is followed by double new line character on no new line character. A single new line character was automatically chosen to be not relevant for sentence boundary detection, as the narratives under scrutiny are heavily fragmented with single new line characters. The statistical co-occurrence features required most resource-intensive computations, as they had to be first obtained from the training corpus before being used as features for machine learning.

The main conclusion of this work is that language technologies can be successfully applied to a variety of medical IR and IE scenarios. However, they have to be adapted to the problem domain and to the specific clinical sub-language. A "one fits all" approach will not work. Different document types (e.g. dermatology notes, pathology reports, discharge summaries) form typical sub-languages, which are further modified by regional and institutional documentation styles. Currently, the most promising scenario for the application of language technologies on clinical narratives is to provide a toolset that can be customized to each clinical use case and document type. To this end it is important to identify the minimal amount of structured information that needs to be added to a narrative in order to optimally address the retrieval or extraction task. An important role is played by supporting terminological resources like vocabularies, thesauri, and ontologies. Although the biomedical domain is supported by numerous large terminology systems, most of them only contain English terms, and their adaptation to other languages is time and resource intensive.

In general, language technology approaches can be divided into two categories: knowledge engineering systems providing rule sets for structural annotations on the one hand, and supervised statistical machine learning approaches on the other hand. The latter ones needs a certain amount of manually annotated data to reach the necessary performance. Gold standard creation is therefore of utmost importance to control the quality of the IE or IR systems, also when building language technology systems and resources in an iterative way. This requires compromises between the complexity of methods and the applicability in real world scenarios. Highest performance values are not satisfactory if the required computational complexity violates real time constraints.

8.1 Outlook

Based on the observations within this thesis the following scenarios are postulated to support future clinical search scenarios in semi-structured data.

Annotation service. For certain sub-domains in a clinical information environment optimized annotation servers will enhance semi-structured data with structured and standardized annotations. This will be done on the fly, in-memory or an enhanced

annotated version of the original data is persistently stored. The additional structured content can then be feed an extended index to support a broad range of semantics-enhanced applications such as disease oriented search scenarios or build a low-level data layer for other analytic tools on the top.

Parallel processing. Processing of millions of documents can be done in a batch process, exploiting distributed computing architectures (e.g. UIMA AS or Apache Hadoop [223]) in combination with NLP techniques. Such batch job will execute within a reasonable amount of time, so that, e.g., secondary use search scenarios can benefit from continuously updated document indexes.

Information overload. With an aging population, the amount of information about a particular patient tends to increase, which creates an information overload that collides with the time constraints of clinicians. Optimized patient-based document search in CISs will become an important use case. Such a patient centric view has to handle increasing amounts of documents that require enhanced visualization and content filtering, combined with user interface tools to support quick navigation. This also applies to future personal record systems like ELGA which will bring together clinical documents from different platforms.

"*That's all I have to say about that*"
Forrest Gump, 1994.

Bibliography

[1] Holzinger A., Stocker C., and Dehmer M. Big complex biomedical data: towards a taxonomy of data. In *E-Business and Telecommunications*, pages 3–18. Springer, 2014.

[2] Talend Open Studio. http://www.talend.com/products/talend-open-studio. [Online; accessed 01.10.2015].

[3] Kreuzthaler M., Daumke P., and Schulz S. Semantic retrieval and navigation in clinical document collections. *EHealth2015–Health Informatics Meets EHealth: Innovative Health Perspectives: Personalized Health*, 212:9–14, 2015.

[4] Patterson O., Igo S., and Hurdle J. F. Automatic acquisition of sublanguage semantic schema: Towards the word sense disambiguation of clinical narratives. In *AMIA Annual Symposium Proceedings*, volume 2010, pages 612–616. American Medical Informatics Association, 2010.

[5] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, pages 137–142, Berlin, 1998. Springer.

[6] Chiticariu L., Li Y., and Reiss F. R. Rule-based information extraction is dead! Long live rule-based information extraction systems! In *EMNLP*, pages 827–832, 2013.

[7] Baharudin B., Lee L. H., and Khan K. A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, 1(1): 4–20, 2010.

[8] Zeng Q. T., Redd D., Rindflesch T., and Nebeker J. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1050. American Medical Informatics Association, 2012.

[9] Arnold C. and Speier W. A topic model of clinical reports. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1031–1032. ACM, 2012.

[10] Park S., Choi D., Lee W., Jung D., Kim M., and Moon I.-C. Disease-medicine topic model for prescription record mining. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 86–93. IEEE, 2014.

[11] Manning C. D., Raghavan P., and Schütze H. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.

[12] Baeza-Yates R., Ribeiro-Neto B., and others . *Modern information retrieval*. Addison-Wesley Reading, MA, 1999.

[13] Harter S. and Hert C. Evaluation of information retrieval systems: Approaches, issues, and methods. *Annual Review of Information Science and Technology (ARIST)*, 32:3–94, 1997.

[14] Buckley C. and Voorhees E. M. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40. ACM, 2000.

[15] Di Eugenio B. and Glass M. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101, 2004.

[16] Buckley C. and Voorhees E. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32. ACM, 2004.

[17] Saracevic T. Evaluation of evaluation in information retrieval. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 138–146. ACM, 1995.

[18] Meystre S. M., Savova G., Kipper-Schuler K., and Hurdle J. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics*, 35:128–144, 2008.

[19] HIPAA Identifiers: Anonymizing Data. http://med.stanford.edu/irt/security/hipaa.html. [Online; accessed 28.09.2015].

[20] Chapman W. W., Chu D., and Dowling J. N. ConText: an algorithm for identifying contextual features from clinical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 81–88. Association for Computational Linguistics, 2007.

[21] Gurulingappa H., Fluck J., Hofmann-Apitius M., and Toldo L. Identification of adverse drug event assertive sentences in medical case reports. In *First International Workshop on Knowledge Discovery and Health Care Management (KD-HCM), European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 16–27, 2011.

[22] Gurulingappa H., Mateen-Rajput A., Toldo L., and others . Extraction of potential adverse drug events from medical case reports. *J Biomed Semantics*, 3(1):15, 2012.

[23] Hripcsak G., Kuperman G. J., Friedman C., and Heitjan D. F. A reliability study for evaluating information extraction from radiology reports. *Journal of the American Medical Informatics Association*, 6(2):143–150, 1999.

[24] Hripcsak G. and Heitjan D. F. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35(2):99–110, 2002.

[25] Hripcsak G. and Wilcox A. Reference standards, judges, and comparison subjects. *Journal of the American Medical Informatics Association*, 9(1):1–15, 2002.

[26] Chapman W. W. and Dowling J. N. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. *Journal of Biomedical Informatics*, 39(2):196–208, 2006.

[27] Rindflesch T. C., Pakhomov S. V., Fiszman M., Kilicoglu H., and Sanchez V. R. Medical facts to support inferencing in natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2005, page 634. American Medical Informatics Association, 2005.

[28] Natural language processing (NLP). http://clinfowiki.org/wiki/index.php/Natural_language_processing_(NLP). [Online; accessed 28.09.2015].

[29] Association for Computational Linguistics. https://www.aclweb.org/. [Online; accessed 28.09.2015].

[30] Association for Computational Linguistics - European Chapter. https://www.aclweb.org/website/eacl. [Online; accessed 28.09.2015].

[31] SIGIR Special Interst Group on Information Retrieval. http://sigir.org/. [Online; accessed 28.09.2015].

[32] The ISCA web - International Speech Communication Associatoin. http://www.isca-speech.org/iscaweb/. [Online; accessed 28.09.2015].

[33] European Association for Mashine Translation. http://www.eamt.org/index.php. [Online; accessed 28.09.2015].

[34] i2b2 - Informatics for Integrating Biology & the Bedside - Data Sets. https://www.aclweb.org/website/eacl. [Online; accessed 28.09.2015].

[35] SemEval-2014 : Semantic Evaluation Exercises. http://alt.qcri.org/semeval2014/, . [Online; accessed 28.09.2015].

[36] SemEval-2015 : Semantic Evaluation Exercises. http://alt.qcri.org/semeval2015/, . [Online; accessed 28.09.2015].

[37] Medical Track. http://trec.nist.gov/data/medical.html, . [Online; accessed 28.09.2015].

[38] Louhi 2015 - the Sixth International Workshop on Health Text Mining and Information Analysis. https://louhi2015.limsi.fr/. [Online; accessed 28.09.2015].

[39] Apache UIMA. https://uima.apache.org/, . [Online; accessed 28.09.2015].

[40] Savova G. K., Masanz J. J., Ogren P. V., Zheng J., Sohn S., Kipper-Schuler K. C., and Chute C. G. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[41] Patterson O., Forbush T., Saini S., Moser S., and DuVall S. Classifying the indication for colonoscopy procedures: A comparison of NLP approaches in a diverse national healthcare system. *Studies in Health Technology and Informatics*, 216:614–618, 2015.

[42] Friedman C., Johnson S. B., Forman B., and Starren J. Architectural requirements for a multipurpose natural language processor in the clinical environment. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 347. American Medical Informatics Association, 1995.

[43] Friedman C., Hripcsak G., DuMouchel W., Johnson S. B., and Clayton P. D. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(01):83–108, 1995.

[44] List of territorial entities where German is an official language. https://en.wikipedia.org/wiki/List_of_territorial_entities_where_German_is_an_official_language, . [Online; accessed 28.09.2015].

[45] Kreuzthaler M., Bloice M., Simonic K.-M., and Holzinger A. Navigating through very large sets of medical records: an information retrieval evaluation architecture for non-standardized text. In *Proceedings of the 7th Conference on Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society: Information Quality in e-Health*, pages 455–470. Springer-Verlag, 2011.

[46] Roberts A., Gaizauskas R., Hepple M., Demetriou G., Guo Y., Setzer A., and Roberts I. Semantic annotation of clinical text: The CLEF corpus. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 19–26, 2008.

[47] Pestian J., Brew C., Matykiewicz P., Hovermale D., Johnson N., Cohen K., and Duch W. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104. Association for Computational Linguistics, 2007.

[48] Hersh W., Müller H., Jensen J., Yang J., Gorman P., and Ruch P. Advancing biomedical image retrieval: development and analysis of a test collection. *Journal of the American Medical Informatics Association*, 13(5):488, 2006. ISSN 1527-974X.

[49] Müller H., Deselaers T., Deserno T., Clough P., Kim E., and Hersh W. Overview of the imageclefmed 2006 medical retrieval and medical annotation tasks. In *Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 595–608. Springer, 2007.

[50] Ogren P., Savova G., Buntrock J., and Chute C. Building and evaluating annotated corpora for medical NLP systems. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1050. American Medical Informatics Association, 2006.

[51] Roberts A., Gaizauskas R., Hepple M., Davis N., Demetriou G., Guo Y., Kola J., Roberts I., Setzer A., Tapuria A., and others . The CLEF corpus: semantic annotation of clinical text. In *AMIA Annual Symposium Proceedings*, volume 2007, page 625. American Medical Informatics Association, 2007.

[52] Uzuner O., Luo Y., and Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5): 550, 2007.

[53] Uzuner O., Goldstein I., Luo Y., and Kohane I. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*, 15(1):14–24, 2008. [PubMed Central:PMC2274873] [DOI:10.1197/jamia.M2408] [PubMed:17947624].

[54] University of Pittsburgh - Department of Biomedical Informatics. http://www.dbmi.pitt.edu/. [Online; accessed 28.09.2015].

[55] Pradhan S., Elhadad N., Chapman W., Manandhar S., and Savova G. Semeval-2014 task 7: Analysis of clinical text. *SemEval 2014*, 199(99):54, 2014.

[56] Elhadad N., Pradhan S., Gorman S. L., Manandhar S., Chapman W., and Savova G. Semeval-2015 task 14: Analysis of clinical text. *Notes*, 298(133):100, 2015.

[57] Bethard S., Derczynski L., Pustejovsky J., and Verhagen M. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics*, 2015.

[58] Hahn U., Romacker M., and Schulz S. MEDSYNDIKATE—a natural language system for the extraction of medical information from findings reports. *International Journal of Medical Informatics*, 67(1):63–74, 2002.

[59] Wermter J. and Hahn U. An annotated german-language medical text corpus as language resource. In *Proceedings of the LREC 2004 Workshop on Building and Evaluating Resources for Biomedical Text Mining*. Citeseer, 2004.

[60] Faessler E., Hellrich J., and Hahn U. Disclose models, hide the data—how to make use of confidential corpora without seeing sensitive raw data. In *LREC 2014 – Proceedings of the 9th Language Resources and Evaluation Conference.*, pages 4230–4237, May 2014.

[61] Geierhofer R. and Holzinger A. Creating an annotated set of medical reports to evaluate information retrieval techniques. In *Proc. of I-MEDIA '07 and I-SEMANTICS '07*, pages 331–337. Citeseer, 2007.

[62] Schmiedberger E. and Errath M. Automatisches Tagging von Pathologie-Befunden. In *Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)*, volume 54. German Medical Science GMS Publishing House, September 2009.

[63] The MedKAT Pipeline. http://ohnlp.sourceforge.net/MedKATp/, . [Online; accessed 28.09.2015].

[64] Cornia R., Patterson O., Ginter T., and Duvall S. Rapid NLP development with leo. In *AMIA Annu Symp Proc*, 2014.

[65] IBM. http://www.ibm.com/us/en/. [Online; accessed 28.09.2015].

[66] Averbis Text Analytics. https://averbis.com/en/. [Online; accessed 28.09.2015].

[67] Jena University Language & Information Engineering Lab. http://www.julielab.de/. [Online; accessed 28.09.2015].

[68] Aronson A. R. Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[69] SemRep - Semantic Knowledge Representation. http://semrep.nlm.nih.gov/, . [Online; accessed 28.09.2015].

[70] Liu Y., Bill R., Fiszman M., Rindflesch T., Pedersen T., Melton G. B., and Pakhomov S. V. Using SemRep to label semantic relations extracted from clinical text. In *AMIA Annual Symposium Proceedings*, volume 2012, page 587. American Medical Informatics Association, 2012.

[71] HITEx - Health Information Text Extraction. https://www.i2b2.org/software/projects/hitex/hitex_manual.html. [Online; accessed 28.09.2015].

[72] Zeng Q. T., Goryachev S., Weiss S., Sordo M., Murphy S. N., and Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 6(1):30, 2006.

[73] Cunningham H. GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254, 2002.

[74] Apache OpenNLP. https://opennlp.apache.org/. [Online; accessed 01.10.2015].

[75] LingPipe. http://alias-i.com/lingpipe/. [Online; accessed 28.09.2015].

[76] MALLET - MAchine Learning for LanguagE Toolkit. http://mallet.cs.umass.edu/. [Online; accessed 28.09.2015].

[77] Hearst M. A., Dumais S., Osman E., Platt J., and Scholkopf B. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28, 1998.

[78] Schölkopf B. and Smola A. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, 2002.

[79] Cristianini N. and Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, 2000.

[80] Cortes C. and Vapnik V. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[81] Bishop C. M. and others . *Pattern Recognition and Machine Learning*, volume 1. Springer New York, New York, 2006.

[82] Salton G., Wong A., and Yang C. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):620, 1975.

[83] Boerjesson E. and Hofsten C. A vector model for perceived object rotation and translation in space. *Psychological Research*, 38(2):209–230, 1975.

[84] Abdou S. and Savoy J. Searching in medline: Query expansion and manual indexing evaluation. *Information Processing & Management*, 44(2):781–789, 2008.

[85] Amati G. and Van Rijsbergen C. J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[86] The Thirteenth Text Retrieval Conference (TREC 2004). http://trec.nist.gov/pubs/trec13/t13_proceedings.html. [Online; accessed 28.09.2015].

[87] Landauer T. and Dumais S. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[88] Landauer T., Foltz P., and Laham D. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[89] Deerwester S. C., Dumais S. T., Landauer T. K., Furnas G. W., and Harshman R. A. Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407, 1990.

[90] Kleene S. C., Bruijn de N., Groot de J., and Zaanen A. C. *Introduction to Metamathematics*. van Nostrand New York, 1952.

[91] McNaughton R. and Yamada H. Regular expressions and state graphs for automata. *IEEE Transactions on Electronic Computers*, 1(1):39–47, 1960.

[92] Thompson K. Regular expression search algorithms. *CACM*, 11(6):419–422, 1968.

[93] Lam M., Sethi R., Ullman J., and Aho A. *Compilers: Principles, Techniques, and Tools*. Pearson Education, Inc, 2006.

[94] Sipser M. *Introduction to the Theory of Computation*. Cengage Learning, 2012.

[95] Class Pattern. http://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html. [Online; accessed 28.09.2015].

[96] JFLAP. http://www.jflap.org/. [Online; accessed 28.09.2015].

[97] Markó K., Schulz S., and Hahn U. Morphosaurus design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods Inf Med*, 44(4):537–545, 2005.

[98] Daumke P. *Das MorphoSaurus-System Lösungen für die linguistischen Herausforderungen des Information Retrievals in der Medizin*. PhD thesis, Universitätsbibliothek Freiburg, 2007.

[99] Hall A. and Walton G. Information overload within the health care system: a literature review. *Health Information & Libraries Journal*, 21(2):102–108, 2004.

[100] Holzinger A., Geierhofer R., and Errath M. Semantische Informationsextraktion in medizinischen Informationssystemen. *Informatik-Spektrum*, 30(2):69–78, 2007.

[101] Geierhofer R. and Errath M. AURAWeb: Kostengünstige Einbindung von Legacy-Daten in ein KIS. In *Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie*, volume 51 of *Abstractband*, September 2006.

[102] Wilcox A. and Hripcsak G. Classification algorithms applied to narrative reports. In *Proceedings of the AMIA Symposium*, page 455. American Medical Informatics Association, 1999.

[103] McCowan I., Moore D., and Fry M.-J. Classification of cancer stage from free-text histology reports. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 5153–5156. IEEE, 2006.

[104] McCowan I., Moore D., and Fry M.-J. Automated Cancer Stage Classification from Free-text Histology Reports. *HIC 2006 and HINZ 2006: Proceedings*, 1:214, 2006.

[105] Nguyen A., Moore D., McCowan I., and Courage M.-J. Multi-class classification of cancer stages from free-text histology reports using support vector machines. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 5140–5143. IEEE, 2007.

[106] Jouhet V., Defossez G., Burgun A., Le Beux P., Levillain P., Ingrand P., Claveau V., and others . Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods of Information in Medicine*, 51(3):242, 2012.

[107] Hiissa M., Pahikkala T., Suominen H., Lehtikunnas T., Back B., Karsten H., Salanterä S., and Salakoski T. Towards automated classification of intensive care nursing narratives. *International Journal of Medical Informatics*, 76:S362–S368, 2007.

[108] Spat S., Cadonna B., Rakovac I., Gutl C., Leitner H., Stark G., Beck P., and others . Multi-label text classification of german language medical documents. *Stud Health Technol Inform*, 129:1460–1461, 2007.

[109] Spat S., Cadonna B., Rakovac I., Gutl C., Leitner H., Stark G., and Beck P. Enhanced information retrieval from narrative german-language clinical text documents using automated document classification. *Stud Health Technol Inform*, 136:473–478, 2008.

[110] Shiner B., D'Avolio L. W., Nguyen T. M., Zayed M. H., Watts B. V., and Fiore L. Automated classification of psychotherapy note text: implications for quality assessment in PTSD care. *Journal of evaluation in clinical practice*, 18(3):698–701, 2012.

[111] Dieperink M., Erbes C., Leskela J., Kaloupek D., and others . Comparison of treatment for post-traumatic stress disorder among three department of veterans affairs medical centers. *Military Medicine*, 170(4):305, 2005.

[112] Afzal Z., Schuemie M. J., Blijderveen van J. C., Sen E. F., Sturkenboom M. C., and Kors J. A. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. *BMC Medical Informatics and Decision Making*, 13(1):30, 2013.

[113] Hofmann T. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57. ACM, 1999.

[114] Blei D., Ng A., and Jordan M. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[115] Mao W. and Chu W. Free-text medical document retrieval via phrase-based vector space model. In *Proceedings of the AMIA Symposium*, page 489. American Medical Informatics Association, 2002.

[116] Hliaoutakis A., Varelas G., Voutsakis E., Petrakis E., and Milios E. Information retrieval by semantic similarity. *International Journal on Semantic Web & Information Systems*, 2(3):55–73, July-September 2006.

[117] Liu Z. and Chu W. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval*, 10(2):173–202, 2007.

[118] Fautsch C. and Savoy J. Adapting the tf-idf vector-space model to domain specific information retrieval. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1708–1712. ACM, 2010.

[119] Chute C. G., Yang Y., and Evans D. Latent semantic indexing of medical diagnoses using UMLS semantic structures. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 185. American Medical Informatics Association, 1991.

[120] Evans D. A., Ginther-Webster K., Hart M., Lefferts R. G., and Monarch I. Automatic indexing using selective NLP and first-order thesauri. In *RIAO*, volume 91, pages 624–643, 1991.

[121] Chute C. G. and Yang Y. An evaluation of concept based latent semantic indexing for clinical information retrieval. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 639. American Medical Informatics Association, 1992.

[122] Kintsch W. The potential of latent semantic analysis for machine grading of clinical case summaries. *Journal of Biomedical Informatics*, 35(1):3–7, 2002.

[123] Cohen A. and Hersh W. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57, 2005.

[124] Cohen T., Blatter B., and Patel V. Simulating expert clinical comprehension: Adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative. *Journal of Biomedical Informatics*, 41(6):1070–1087, 2008.

[125] Association A. P. Diagnostic and statistical manual of mental disorders (DSM). *Washington, DC: American Psychiatric Association*, 1:143–147, 1994.

[126] Widdows D. and Peters S. Word vectors and quantum logic: Experiments with negation and disjunction. *Mathematics of Language*, 8:141–154, 2003.

[127] Ginter F., Suominen H., Pyysalo S., and Salakoski T. Combining hidden markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. *International Journal of Medical Informatics*, 78(12):e1–e6, 2009.

[128] Tremblay M. C., Berndt D. J., Luther S. L., Foulis P. R., and French D. D. Identifying fall-related injuries: Text mining the electronic medical record. *Information Technology and Management*, 10(4):253–265, 2009.

[129] Elvevåg B., Foltz P. W., Weinberger D. R., and Goldberg T. E. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93(1):304–316, 2007.

[130] Henriksson A., Hassel M., and Kvist M. Diagnosis code assignment support using random indexing of patient records–a qualitative feasibility study. In *Artificial Intelligence in Medicine*, pages 348–352. Springer, 2011.

[131] Henriksson A. and Hassel M. Election of diagnosis codes-words as responsible citizens. In *Proceedings of Louhi 2011-3rd International Workshop on Health Document Text Mining and Information Analysis*, 2011.

[132] Henriksson A. and Hassel M. Exploiting structured data, negation detection and SNOMED CT terms in a random indexing approach to clinical coding. In *Proceedings of Workshop on Biomedical Natural Language Processing*, pages 3–10, 2011.

[133] Henriksson A. and Hassel M. Optimizing the dimensionality of clinical term spaces for improved diagnosis coding support. In *4th International Louhi Workshop on Health Document Text Mining and Information Analysis Sydney, NSW, Australia, 11-12 February 2013*. NICTA, 2013.

[134] Henriksson A., Moen H., Skeppstedt M., Daudaravičius V., and Duneld M. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5(1):1–25, 2014.

[135] Henriksson A., Conway M., Duneld M., and Chapman W. W. Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. In *AMIA Annual Symposium Proceedings*, volume 2013, page 600. American Medical Informatics Association, 2013.

[136] Henriksson A., Skeppstedt M., Kvist M., Duneld M., and Conway M. Corpus-driven terminology development: Populating swedish SNOMED CT with synonyms extracted from electronic health records. In *2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), Sofia, Bulgaria, August 4-9, 2013*, pages 36–44. Association for Computational Linguistics, 2013.

[137] Henriksson A., Moen H., Skeppstedt M., Eklund A.-M., Daudaravičius V., and Hassel M. Synonym extraction of medical terms from clinical text using combinations of word space models. In *5th International Symposium on Semantic Mining in Biomedicine (SMBM), September 3-4 2012, Zurich, Switzerland*, pages 10–17, 2012.

[138] Henriksson A., Kvist M., Hassel M., and Dalianis H. Exploration of adverse drug reactions in semantic vector space models of clinical text. In *Proceedings of ICML 2012-The 29th International Conference on Machine Learning*, 2012.

[139] Henriksson A., Dalianis H., and Kowalski S. Generating features for named entity recognition by learning prototypes in semantic space: The case of de-identifying health records. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 450–457. IEEE, 2014.

[140] Moen H., Ginter F., Marsi E., Peltonen L.-M., Salakoski T., and Salanterä S. Care episode retrieval: distributional semantic models for information retrieval in the clinical domain. *BMC Medical Informatics and Decision Making*, 15(Suppl 2):S2, 2015.

[141] Zhang S. and Elhadad N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, 46 (6):1088–1098, 2013.

[142] Natarajan K. *Analysis of Search on Clinical Narrative within the EHR.* PhD thesis, Columbia University, 2012.

[143] Cohen T. and Widdows D. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390–405, 2009.

[144] Bradford R. B. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 153–162. ACM, 2008.

[145] Voorhees E. and Tong R. Overview of the trec 2012 medical records track. In *Proc. of TREC*, volume 4, 2012.

[146] Edinger T., Cohen A. M., Bedrick S., Ambert K., and Hersh W. Barriers to retrieving patient information from electronic health record data: Failure analysis from the trec medical records track. In *AMIA Annual Symposium Proceedings*, volume 2012, page 180. American Medical Informatics Association, 2012.

[147] Rehurek R. Fast and faster: A comparison of two streamed matrix decomposition algorithms. In *NIPS2010 Workshop on Low-rank Methods for Large-scale Machine Learning*, 2010.

[148] Rehurek R. Subspace tracking for latent semantic analysis. In *Advances in Information Retrieval*, pages 289–300. Springer, 2011.

[149] Vigna S. Distributed, large-scale latent semantic analysis by index interpolation. In *Proceedings of the 3rd International Conference on Scalable Information Systems*, page 18. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008.

[150] Cavanagh J. M., Potok T. E., and Cui X. Parallel latent semantic analysis using a graphics processing unit. In *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers*, pages 2505–2510. ACM, 2009.

[151] Sahlgren M. An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, volume 5, 2005.

[152] Chapman W. W., Bridewell W., Hanbury P., Cooper G. F., and Buchanan B. G. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*, 34(5):301–310, Oct 2001. [DOI:10.1006/jbin.2001.1029] [PubMed:12123149].

[153] Litton J. Biobank informatics: connecting genotypes and phenotypes. *Methods in Molecular Biology*, 675:343–361, 2011. [DOI:10.1007/978-1-59745-423-0_21] [PubMed:20949402].

[154] Eder J., Dabringer C., Schicho M., and Stark K. Information systems for federated biobanks. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems I*, pages 156–190. Springer, 2009.

[155] Kreuzthaler M., Schulz S., and Berghold A. Secondary use of electronic health records for building cohort studies through top-down information extraction. *Journal of Biomedical Informatics*, 53:188–195, 2015.

[156] Karolinska Mammography Project for Risk Prediction of Breast Cancer. http://karmastudy.org/. [Online; accessed 01.10.2015].

[157] Uzuner O., South B. R., Shen S., and DuVall S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–556, 2011. [PubMed Central:PMC3168320] [DOI:10.1136/amiajnl-2011-000203] [PubMed:21685143].

[158] Bruijn de B., Cherry C., Kiritchenko S., Martin J., and Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc*, 18(5):557–562, 2011. [PubMed Central:PMC3168309] [DOI:10.1136/amiajnl-2011-000150] [PubMed:21565856].

[159] Heintzelman N. H., Taylor R. J., Simonsen L., Lustig R., Anderko D., Haythornthwaite J. A., Childs L. C., and Bova G. S. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. *J Am Med Inform Assoc*, 20(5):898–905, 2013. [PubMed Central:PMC3756253] [DOI:10.1136/amiajnl-2012-001076] [PubMed:23144336].

[160] Skeppstedt M., Kvist M., Nilsson G. H., and Dalianis H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. *J Biomed Inform*, 49:148–158, Jun 2014. [DOI:10.1016/j.jbi.2014.01.012] [PubMed:24508177].

[161] Botsis T., Hartvigsen G., Chen F., and Weng C. Secondary use of EHR: Data quality issues and informatics opportunities. *AMIA Jt Summits Transl Sci Proc*, 2010:1–5, 2010. [PubMed Central:PMC3041534] [PubMed:21347133].

[162] Antolik J. Automatic annotation of medical records. *Stud Health Technol Inform*, 116:817–822, 2005. [PubMed:16160359].

[163] Ciravegna F. Adaptive information extraction from text by rule induction and generalisation. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence-Volume 2*, pages 1251–1256. Morgan Kaufmann Publishers Inc., 2001.

[164] Roque F. S., Jensen P. B., Schmock H., Dalgaard M., Andreatta M., Hansen T., Søby K., Bredkjær S., Juul A., Werge T., Jensen L. J., and Brunak S. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput. Biol.*, 7(8):e1002141, Aug 2011. [PubMed Central:PMC3161904] [DOI:10.1371/journal.pcbi.1002141] [PubMed:21901084].

[165] OMIM - Online Mendelian Inheritance in Man. http://www.omim.org/. [Online; accessed 01.10.2015].

[166] Martinell M., Stalhammar J., and Hallqvist J. Automated data extraction–a feasible way to construct patient registers of primary care utilization. *Ups. J. Med. Sci.*, 117(1):52–56, Mar 2012. [PubMed Central:PMC3282243] [DOI:10.3109/03009734.2011.653015] [PubMed:22335391].

[167] Xu H., Fu Z., Shah A., Chen Y., Peterson N. B., Chen Q., Mani S., Levy M. A., Dai Q., and Denny J. C. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA Annu Symp Proc*, 2011: 1564–1572, 2011. [PubMed Central:PMC3243156] [PubMed:22195222].

[168] Segagni D., Tibollo V., Dagliati A., Perinati L., Zambelli A., Priori S., and Bellazzi R. The ONCO-i2b2 project: integrating biobank information and clinical data to support translational research in oncology. *Stud Health Technol Inform*, 169:887–891, 2011. [PubMed:21893874].

[169] Murphy S. N., Weber G., Mendis M., Gainer V., Chueh H. C., Churchill S., and Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*, 17(2):124–130, 2010. [PubMed Central:PMC3000779] [DOI:10.1136/jamia.2009.000893] [PubMed:20190053].

[170] GATE - General Architecture for Text Engineering. https://gate.ac.uk/. [Online; accessed 01.10.2015].

[171] Cunningham H., Tablan V., Roberts A., and Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Comput. Biol.*, 9(2):e1002854, 2013. [PubMed Central:PMC3567135] [DOI:10.1371/journal.pcbi.1002854] [PubMed:23408875].

[172] UIMA Tutorial and Developers' Guides. http://uima.apache.org/downloads/releaseDocs/2.3.0-incubating/docs/html/tutorials_and_users_guides/tutorials_and_users_guides.html, . [Online; accessed 01.10.2015].

[173] iText - Programmable PDF Software. http://itextpdf.com/. [Online; accessed 01.10.2015].

[174] Hibernate - Everything data. http://hibernate.org/. [Online; accessed 01.10.2015].

[175] Savova G., Kipper-Schuler K., Buntrock J., and Chute C. Uima-based clinical information extraction system. *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP*, 39:39–42, 2008.

[176] Ramakrishnan N., Hanauer D., and Keller B. Mining electronic health records. *Computer*, 43(10):77–81, 2010. [DOI:10.1109/MC.2010.292].

[177] Chapman W. W., Nadkarni P. M., Hirschman L., D'Avolio L. W., Savova G. K., and Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*, 18(5):540–543, 2011. [PubMed Central:PMC3168329] [DOI:10.1136/amiajnl-2011-000465] [PubMed:21846785].

[178] BBMRI-ERIC Biobanking and BioMolecular resources Research Infrastructure. http://bbmri-eric.eu/. [Online; accessed 01.10.2015].

[179] eMERGE network - Electronic Medical Records and Genomics. http://bbmri-eric.eu/. [Online; accessed 01.10.2015].

[180] HeidelTime - a multilingual, cross-domain temporal tagger. https://code.google.com/p/heideltime/. [Online; accessed 01.10.2015].

[181] Gostev M., Fernandez-Banet J., Rung J., Dietrich J., Prokopenko I., Ripatti S., McCarthy M. I., Brazma A., and Krestyaninova M. Sail–a software system for sample and phenotype availability across biobanks and cohorts. *Bioinformatics*, 27(4):589–591, Feb 2011. [PubMed Central:PMC3035801] [DOI:10.1093/bioinformatics/btq693] [PubMed:21169373].

[182] Hripcsak G. and Albers D. J. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*, 20(1):117–121, Jan 2013. [PubMed Central:PMC3555337] [DOI:10.1136/amiajnl-2012-001145] [PubMed:22955496].

[183] Kreuzthaler M. and Schulz S. Detection of sentence boundaries and abbreviations in clinical narratives. *BMC Medical Informatics and Decision Making*, 15(Suppl 2): S4, 2015.

[184] Xu H., Stetson P., and Friedman C. A study of abbreviations in clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2007, pages 821–825, 2007.

[185] Wiesenauer M., Johner C., and Röhrig R. Secondary use of clinical data in healthcare providers–an overview on research, regulatory and ethical requirements. *Studies in Health Technology and Informatics*, 180:614–618, 2012.

[186] International Classification of Diseases. http://www.who.int/classifications/icd/en/. [Online; accessed 01.10.2015].

[187] Xu H., Stetson P., and Friedman C. Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. In *AMIA Annual Symposium Proceedings*, volume 2012, pages 1004–1013, 2012.

[188] Okazaki N., Ananiadou S., and Tsujii J. Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26(9):1246–1253, 2010.

[189] Suominen H., Salanterä S., Velupillai S., Chapman W. W., Savova G., Elhadad N., Pradhan S., South B. R., Mowery D. L., Jones G. J., and others . Overview of the ShARe/CLEF ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231, 2013.

[190] Unified Medical Language System. http://www.nlm.nih.gov/research/umls/. [Online; accessed 01.10.2015].

[191] Patrick J., Safari L., and Ou Y. ShaARe/CLEF eHealth 2013 normalization of acronyms/abbreviation challenge. In *CLEF 2013 Evaluation Labs and Workshop Abstracts - Working Notes*, 2013.

[192] Friedman C., Alderson P. O., Austin J. H., Cimino J. J., and Johnson S. B. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

[193] Friedman C., Hripcsak G., Shagina L., and Liu H. Representing information in patient reports using natural language processing and the extensible markup language. *Journal of the American Medical Informatics Association*, 6(1):76–87, 1999.

[194] Wu Y., Rosenbloom S., Denny J., Miller A., Mani S., DA G., and Xu H. Detecting abbreviations in discharge summaries using machine learning methods. In *AMIA Annual Symposium Proceedings*, volume 2011, pages 1541–1549, 2011.

[195] Wu Y., Denny J., Rosenbloom S., Miller R., Giuse D., and Xu H. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In *AMIA Annual Symposium Proceedings*, volume 2012, pages 997–1003, 2012.

[196] Wu Y., Denny J., Rosenbloom S., Miller R. A., Giuse D. A., Song M., and Xu H. A prototype application for real-time recognition and disambiguation of clinical abbreviations. In *Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Informatics*, pages 7–8, 2013.

[197] Kreuzthaler M. and Schulz S. Disambiguation of period characters in clinical narratives. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@EACL*, pages 96–100, 2014.

[198] Gillick D. Sentence boundary detection and the problem with the U.S. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 241–244. Association for Computational Linguistics, 2009.

[199] Kiss T. and Strunk J. Scaled log likelihood ratios for the detection of abbreviations in text corpora. In *Proceedings of the 19th International Conference on Computational Linguistics – Volume 2*, pages 1–5. Association for Computational Linguistics, 2002.

[200] Apache Lucene Core. http://lucene.apache.org/core/. [Online; accessed 01.10.2015].

[201] Hagerup T. and Rüb C. A guided tour of chernoff bounds. *Information Processing Letters*, 33(6):305–308, 1990.

[202] O'Donnell R. Probability and Computing (CMU course 15-359) Lecture Notes, Lecture 10. Carnegie Mellon Univerity, School of Computer Science, 2009. URL http://www.cs.cmu.edu/%7Eodonnell/papers/probability-and-computing-lecture-notes.pdf.

[203] Free German Dictionary. http://sourceforge.net/projects/germandict/, . [Online; accessed 01.10.2015].

[204] Pschyrembel . Klinisches Wörterbuch. CD-ROM Version 1/97, 1997.

[205] Netdoktor. http://www.netdoktor.at/. [Online; accessed 01.10.2015].

[206] Medizinische Abkürzungen. http://de.wikipedia.org/wiki/Medizinische_Abk%C3%BCrzungen, . [Online; accessed 01.10.2015].

[207] Deutsche Abkürzungen. http://de.wiktionary.org/wiki/Kategorie:Abk%C3%BCrzung_(Deutsch), . [Online; accessed 01.10.2015].

[208] Deutsche Grammatik 2.0. http://www.deutschegrammatik20.de/. [Online; accessed 01.10.2015].

[209] LIBLINEAR – A Library for Large Linear Classification. http://www.csie.ntu.edu.tw/%7Ecjlin/liblinear/. [Online; accessed 01.10.2015].

[210] Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/. [Online; accessed 01.10.2015].

[211] Hsu C.-W., Chang C.-C., Lin C.-J., and others . A Practical Guide to Support Vector Classification, 2010.

[212] Joachims T. *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms*. Kluwer Academic Publishers, Norwell, 2002.

[213] Guyon I., Weston J., Barnhill S., and Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[214] Dunning T. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[215] Surprise and coincidence - musings from the long tail. http://tdunning.blogspot.co.at/2008/03/surprise-and-coincidence.html. [Online; accessed 01.10.2015].

[216] Apache Mahout. https://mahout.apache.org/. [Online; accessed 01.10.2015].

[217] Kiss T. and Strunk J. Viewing sentence boundary detection as collocation identification. In *Proceedings of KONVENS 2002*, pages 75–82, 2002.

[218] Kiss T. and Strunk J. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525, 2006.

[219] Buyko E., Wermter J., Poprat M., and Hahn U. Automatically adapting an NLP core engine to the biology domain. In *Proceedings of the Joint BioLINK-Bio-Ontologies Meeting. A Joint Meeting of the ISMB Special Interest Group on Bio-Ontologies and the BioLINK Special Interest Group on Text Data Mining in Association with ISMB*, pages 65–68, 2006.

[220] Friedman C. A broad-coverage natural language processing system. In *Proceedings of the AMIA Symposium*, pages 270–274. American Medical Informatics Association, 2000.

[221] MedKAT. http://ohnlp.sourceforge.net/MedKATp/#d4e5, . [Online; accessed 01.10.2015].

[222] UIMA Asynchronous Scaleout. https://uima.apache.org/downloads/releaseDocs/2.3.0-incubating/docs-uima-as/html/uima_async_scaleout/uima_async_scaleout.html, . [Online; accessed 26.10.2015].

[223] Apache Hadoop. https://hadoop.apache.org/. [Online; accessed 06.10.2015].

[224] Elsevier - Copyright. http://www.elsevier.com/about/company-information/policies/copyright#authors-rights. [Online; accessed 01.10.2015].

[225] BioMed Central license agreement. http://www.biomedcentral.com/about/license. [Online; accessed 28.09.2015].

[226] Creative Commons Attribution 4.0 International. http://creativecommons.org/licenses/by/4.0/legalcode. [Online; accessed 28.09.2015].

Appendix A

Natural Language Technologies

A.1 Stop word list

aber, alle, allem, allen, aller, alles, als, also, am, an, ander, andere, anderem, anderen, anderer, anderes, anderm, andern, anders, auch, auf, aus, bei, bin, bis, bist, da, damit, dann, der, den, des, dem, die, das, daß, dass, derselbe, derselben, denselben, desselben, demselben, dieselbe, dieselben, dasselbe, dazu, dein, deine, deinem, deinen, deiner, deines, denn, derer, dessen, dich, dir, du, dies, diese, diesem, diesen, dieser, dieses, doch, dort, durch, ein, eine, einem, einen, einer, eines, einig, einige, einigem, einigen, einiger, einiges, einmal, er, ihn, ihm, es, etwas, euer, eure, eurem, euren, eurer, eures, für, fuer, gegen, gewesen, hab, habe, haben, hat, hatte, hatten, hier, hin, hinter, ich, mich, mir, ihr, ihre, ihrem, ihren, ihrer, ihres, euch, im, in, indem, ins, ist, jede, jedem, jeden, jeder, jedes, jene, jenem, jenen, jener, jenes, jetzt, kann, kein, keine, keinem, keinen, keiner, keines, können, koennen, könnte, koennte, machen, man, manche, manchem, manchen, mancher, manches, mein, meine, meinem, meinen, meiner, meines, mit, muss, musste, nach, nicht, nichts, noch, nun, nur, ob, oder, ohne, sehr, sein, seine, seinem, seinen, seiner, seines, selbst, sich, sie, ihnen, sind, so, solche, solchem, solchen, solcher, solches, soll, sollte, sondern, sonst, über, ueber, um, und, uns, unse, unsem, unsen, unser, unses, unter, viel, vom, von, vor, während, waehrend, war, waren, warst, was, weg, weil, weiter, welche, welchem, welchen, welcher, welches, wenn, werde, werden, wie, wieder, will, wir, wird, wirst, wo, wollen, wollte, würde, wuerde, würden, wuerden, zu, zum, zur, zwar, zwischen

A.2 Clinical information retrieval

A.2.1 Results

	Snowball$_{25}^{InB2}$				Morpho$_{15}^{InB2}$				Mixed$_{20}^{InB2}$			
Topic number	MAP	P_{10}	P_{20}	P_{30}	MAP	P_{10}	P_{20}	P_{30}	MAP	P_{10}	P_{20}	P_{30}
1	0.25	0.10	0.30	0.40	0.87	1.00	1.00	1.00	0.83	1.00	0.95	0.97
2	0.01	0.00	0.00	0.00	0.67	0.60	0.65	0.63	0.67	0.50	0.65	0.73
3	0.03	0.10	0.05	0.07	0.34	0.60	0.45	0.47	0.48	0.60	0.50	0.57
4	0.04	0.00	0.00	0.00	0.91	1.00	1.00	1.00	0.90	1.00	0.95	0.90
5	0.13	0.40	0.25	0.23	0.12	0.40	0.25	0.23	0.13	0.20	0.20	0.23
6	0.02	0.00	0.00	0.00	0.71	1.00	1.00	0.97	0.74	1.00	1.00	0.97
7	0.07	0.10	0.05	0.03	0.22	0.40	0.30	0.27	0.24	0.50	0.25	0.17
8	0.02	0.00	0.00	0.00	0.06	0.30	0.15	0.10	0.03	0.00	0.05	0.03
9	0.01	0.00	0.00	0.00	0.42	0.90	0.60	0.57	0.35	0.40	0.60	0.60
10	0.01	0.00	0.00	0.00	0.85	0.90	0.90	0.77	0.85	1.00	1.00	0.73
11	0.38	0.80	0.50	0.43	0.44	1.00	0.95	0.83	0.44	1.00	0.85	0.80
12	0.09	0.30	0.15	0.17	0.35	0.50	0.50	0.43	0.28	0.20	0.30	0.27
13	0.02	0.00	0.00	0.00	0.44	0.60	0.75	0.80	0.40	0.60	0.80	0.83
14	0.02	0.00	0.00	0.00	0.26	0.60	0.40	0.47	0.15	0.20	0.35	0.30
15	0.15	0.20	0.15	0.23	0.73	0.90	0.85	0.90	0.46	0.50	0.55	0.57
16	0.02	0.00	0.00	0.00	0.98	1.00	1.00	1.00	0.98	1.00	1.00	1.00
17	0.26	0.50	0.50	0.53	0.75	0.80	0.90	0.90	0.74	0.90	0.90	0.93
18	0.01	0.00	0.00	0.00	0.16	0.10	0.10	0.17	0.15	0.20	0.15	0.13
19	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00
20	0.15	0.10	0.10	0.10	0.65	0.40	0.60	0.67	0.53	0.20	0.50	0.63
21	0.10	0.00	0.00	0.03	0.59	0.40	0.55	0.60	0.44	0.10	0.40	0.50
22	0.07	0.00	0.00	0.00	0.11	0.00	0.05	0.07	0.10	0.10	0.05	0.03
23	0.27	0.40	0.30	0.27	0.50	0.90	0.60	0.60	0.51	1.00	0.90	0.80
24	0.55	0.90	0.95	0.93	0.86	1.00	1.00	1.00	0.86	1.00	1.00	0.97
25	0.74	1.00	0.95	0.97	0.77	0.90	0.95	0.93	0.79	0.80	0.90	0.93
26	0.06	0.00	0.00	0.00	0.47	0.90	0.80	0.87	0.53	1.00	0.90	0.83
Mean	0.13	0.19	0.16	0.17	0.51	0.66	0.63	0.62	0.48	0.58	0.60	0.59

Table A.1: Detailed information need evaluation for the language register *Layperson* at their local maxima MAP$_{max}$ (Figure 5.1(a)) depending on the degree of dimension reduction per preprocessing step (Snowball, Morpho, Mixed) together with their Precision at 10, 20 and 30 values (P_{10}, P_{20}, P_{30}) for the weighting scheme I(n)B2.

	Snowball$_{55}^{InB2}$				Morpho$_{15}^{InB2}$				Mixed$_{20}^{InB2}$			
Topic number	MAP	P_{10}	P_{20}	P_{30}	MAP	P_{10}	P_{20}	P_{30}	MAP	P_{10}	P_{20}	P_{30}
1	0.82	1.00	1.00	1.00	0.87	1.00	1.00	1.00	0.83	1.00	0.95	0.97
2	0.68	0.70	0.80	0.80	0.67	0.60	0.65	0.63	0.67	0.50	0.65	0.73
3	0.04	0.10	0.05	0.03	0.34	0.60	0.45	0.47	0.48	0.60	0.50	0.57
4	0.04	0.00	0.00	0.00	0.91	1.00	1.00	1.00	0.90	1.00	0.95	0.90
5	0.02	0.00	0.00	0.00	0.12	0.40	0.25	0.23	0.16	0.30	0.35	0.30
6	0.02	0.00	0.00	0.00	0.71	1.00	1.00	0.97	0.74	1.00	1.00	0.97
7	0.03	0.00	0.05	0.03	0.22	0.40	0.30	0.27	0.24	0.50	0.25	0.17
8	0.02	0.00	0.00	0.03	0.06	0.30	0.15	0.10	0.03	0.00	0.05	0.03
9	0.66	1.00	1.00	0.93	0.42	0.90	0.60	0.57	0.35	0.40	0.60	0.60
10	0.01	0.00	0.00	0.00	0.85	0.90	0.90	0.77	0.85	1.00	1.00	0.73
11	0.57	0.60	0.55	0.57	0.44	1.00	0.95	0.83	0.44	1.00	0.85	0.80
12	0.09	0.30	0.15	0.17	0.35	0.50	0.50	0.43	0.28	0.20	0.30	0.27
13	0.45	1.00	0.95	0.80	0.44	0.60	0.75	0.80	0.41	0.70	0.80	0.83
14	0.02	0.00	0.00	0.00	0.26	0.60	0.40	0.47	0.15	0.20	0.35	0.30
15	0.08	0.20	0.25	0.37	0.73	0.90	0.85	0.90	0.46	0.50	0.55	0.57
16	0.02	0.00	0.00	0.00	0.98	1.00	1.00	1.00	0.98	1.00	1.00	1.00
17	0.18	0.50	0.65	0.60	0.75	0.80	0.90	0.90	0.74	0.90	0.90	0.93
18	0.55	1.00	0.85	0.57	0.39	0.90	0.55	0.40	0.40	0.80	0.45	0.37
19	0.01	0.00	0.00	0.00	0.57	0.80	0.80	0.73	0.27	0.30	0.15	0.13
20	0.19	0.00	0.00	0.00	0.65	0.40	0.60	0.67	0.52	0.20	0.45	0.60
21	0.10	0.00	0.00	0.03	0.59	0.40	0.55	0.60	0.44	0.10	0.40	0.50
22	0.73	1.00	1.00	0.97	0.32	0.20	0.35	0.27	0.31	0.20	0.40	0.27
23	0.27	0.40	0.30	0.27	0.50	0.90	0.60	0.60	0.51	1.00	0.90	0.80
24	0.59	0.70	0.70	0.73	0.86	1.00	1.00	1.00	0.86	1.00	1.00	0.97
25	0.65	1.00	0.95	0.93	0.77	0.90	0.95	0.93	0.79	0.80	0.90	0.93
26	0.02	0.00	0.00	0.00	0.47	0.90	0.80	0.87	0.54	1.00	0.90	0.87
Mean	0.26	0.37	0.36	0.34	0.55	0.73	0.69	0.67	0.51	0.62	0.64	0.62

Table A.2: Detailed information need evaluation for the language register *Netdoktor* at their local maxima MAP$_{max}$ (Figure 5.1(b)) depending on the degree of dimension reduction per preprocessing step (Snowball, Morpho, Mixed) together with their Precision at 10, 20 and 30 values (P_{10}, P_{20}, P_{30}) for the weighting scheme I(n)B2.

	Snowball$_{55}^{InB2}$				Morpho$_{15}^{InB2}$				Mixed$_{15}^{InB2}$			
Topic number	MAP	P_{10}	P_{20}	P_{30}	MAP	P_{10}	P_{20}	P_{30}	MAP	P_{10}	P_{20}	P_{30}
1	0.82	1.00	1.00	1.00	0.87	1.00	1.00	1.00	0.92	1.00	1.00	1.00
2	0.68	0.70	0.80	0.80	0.67	0.60	0.65	0.63	0.71	0.70	0.75	0.73
3	0.08	0.60	0.30	0.23	0.34	0.60	0.45	0.47	0.64	0.90	0.65	0.67
4	0.73	1.00	1.00	1.00	0.91	1.00	1.00	1.00	0.87	0.70	0.85	0.90
5	0.02	0.00	0.00	0.00	0.12	0.40	0.25	0.23	0.15	0.50	0.35	0.23
6	0.55	0.70	0.80	0.83	0.71	1.00	1.00	0.97	0.64	1.00	1.00	0.90
7	0.02	0.10	0.05	0.03	0.33	0.70	0.50	0.40	0.28	0.60	0.45	0.40
8	0.04	0.10	0.05	0.03	0.06	0.30	0.15	0.10	0.03	0.00	0.05	0.03
9	0.66	1.00	1.00	0.93	0.42	0.90	0.60	0.57	0.36	0.40	0.60	0.60
10	0.80	0.80	0.75	0.73	0.78	0.90	0.75	0.70	0.71	0.90	0.70	0.73
11	0.54	0.70	0.55	0.53	0.44	1.00	0.95	0.83	0.41	1.00	0.80	0.73
12	0.23	0.80	0.65	0.67	0.49	0.30	0.35	0.37	0.46	0.50	0.50	0.60
13	0.02	0.00	0.00	0.00	0.56	0.70	0.80	0.83	0.46	0.70	0.70	0.70
14	0.48	0.90	0.90	0.73	0.23	0.50	0.35	0.33	0.16	0.20	0.30	0.23
15	0.09	0.30	0.45	0.43	0.73	0.90	0.85	0.90	0.70	0.90	0.80	0.80
16	0.80	0.70	0.85	0.87	0.98	1.00	1.00	1.00	0.97	1.00	1.00	1.00
17	0.18	0.50	0.65	0.60	0.75	0.80	0.90	0.90	0.78	0.90	0.95	0.93
18	0.55	1.00	0.85	0.57	0.39	0.90	0.55	0.40	0.41	0.90	0.55	0.43
19	0.01	0.00	0.00	0.00	0.57	0.80	0.80	0.73	0.12	0.10	0.05	0.03
20	0.15	0.10	0.10	0.10	0.65	0.40	0.60	0.67	0.67	0.40	0.70	0.77
21	0.41	1.00	0.95	0.90	0.64	1.00	1.00	0.93	0.64	0.90	0.95	0.97
22	0.73	1.00	1.00	0.97	0.32	0.20	0.35	0.27	0.33	0.20	0.15	0.20
23	0.40	0.90	0.95	0.93	0.69	0.70	0.75	0.80	0.73	0.70	0.85	0.90
24	0.59	0.70	0.70	0.73	0.86	1.00	1.00	1.00	0.83	1.00	1.00	0.97
25	0.39	1.00	0.85	0.73	0.12	0.10	0.10	0.17	0.12	0.20	0.15	0.17
26	0.61	0.90	0.85	0.90	0.49	1.00	0.80	0.87	0.21	0.50	0.35	0.30
Mean	0.41	0.63	0.62	0.59	0.54	0.72	0.67	0.66	0.51	0.65	0.62	0.61

Table A.3: Detailed information need evaluation for the language register *Expert* at their local maxima MAP$_{max}$ (Fiure 5.1(c)) depending on the degree of dimension reduction per preprocessing step (Snowball, Morpho, Mixed) together with their Precision at 10, 20 and 30 values (P_{10}, P_{20}, P_{30}) for the weighting scheme I(n)B2.

Appendix B

Copy Right Statements

B.1 Elsevier

Contents of the publication:

> **Kreuzthaler, M.**, Schulz, S., & Berghold, A. (2015). Secondary use of electronic health records for building cohort studies through top-down information extraction. Journal of Biomedical Informatics, 53, (pp. 188-195).
> DOI: http://dx.doi.org/10.1016/j.jbi.2014.10.010

were re-used in Chapter 6 and Chapter 8 according to the Elsevier Coypright - Journal author rights [224] - Statement for personal use:

"Authors can use their articles, in full or in part, for a wide range of scholarly, non-commercial purposes as outlined below:

Use by an author in the author's classroom teaching (including distribution of copies, paper or electronic); Distribution of copies (including through e-mail) to known research colleagues for their personal use (but not for Commercial Use); **Inclusion in a thesis or dissertation (provided that this is not to be published commercially)**; Use in a subsequent compilation of the author's works; Extending the Article to book-length form; Preparation of other derivative works (but not for Commercial Use); Otherwise using or re-using portions or excerpts in other works.

These rights apply for all Elsevier authors who publish their article as either a subscription article or an open access article. In all cases we require that all Elsevier authors always include a full acknowledgement and, if appropriate, a link to the final published version hosted on Science Direct."

Univ.-Prof. Dr.med. Stefan Schulz and Univ.-Prof. Dipl.-Ing. Dr. techn. Andrea Berghold fully acknowledge the inclusion of the publication as part of this thesis.

B.2 BioMed Central

Contents of the publication:

> **Kreuzthaler, M.**, & Schulz, S. (2015). Detection of sentence boundaries and abbreviations in clinical narratives. BMC Medical Informatics and Decision Making, 15 (Suppl 2), S4.
> DOI: 10.1186/1472-6947-15-S2-S4

were re-used for Section 3.1, Chapter 7 and Chapter 8 according to the BioMed Central license agreement [225] and the Creative Commons Attribution License 4.0 [226] which state:

"You are free to: **Share - copy and redistribute the material in any medium or format**; **Adapt - remix, transform, and build upon the material for any purpose, even commercially.** The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms: Attribution - You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. No additional restrictions - You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits."

For inclusion in this thesis minor changes were applied according to feedback of the dissertation committee.

Univ.-Prof. Dr.med. Stefan Schulz fully acknowledges the inclusion of the publication as part of this thesis.

Graz, 28th October 2015 Markus Kreuzthaler

Appendix C

List of Publications

Miñarro-Giménez, J. A., **Kreuzthaler, M.**, & Schulz, S. (2015). Knowledge extraction from MEDLINE by combining clustering with natural language processing. *To appear* in AMIA Annual Symposium Proceedings (Vol. 2015). American Medical Informatics Association.

Miñarro-Giménez, J. A., **Kreuzthaler, M.**, Bernhardt-Melischnig, J., Martínez-Costa, C., & Schulz, S. (2015). Acquiring plausible predications from MEDLINE by clustering MeSH annotations. Studies in Health Technology and Informatics, 216, 716-720.

Kreuzthaler, M., Schulz, S., & Berghold, A. (2015). Secondary use of electronic health records for building cohort studies through top-down information extraction. Journal of Biomedical Informatics, 53, (pp. 188-195).

Kreuzthaler, M., & Schulz, S. (2015). Detection of sentence boundaries and abbreviations in clinical narratives. BMC Medical Informatics and Decision Making, 15 (Suppl 2), S4.

Kreuzthaler, M., Daumke, P., & Schulz, S. (2015). Semantic retrieval and navigation in clinical document collections. eHealth2015 - Health Informatics meets eHealth: Innovative Health Perspectives: Personalized Health, 212, (pp. 9-14).

Schulz, S., Costa, C. M., **Kreuzthaler, M.**, Miñarro-Giménez, J. A., Andersen, U., Jensen, A. B., & Maegaard, B. (2014, May) Semantic relation discovery by using co-occurrence information. In Proceedings of the 4th Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM)@LREC.

Kreuzthaler, M., & Schulz, S. (2014, April). Disambiguation of period characters in clinical narratives. In Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@EACL (pp. 96-100).

Schulz, S., Bernhardt-Melischnig, J., **Kreuzthaler, M.**, Daumke, P., & Boeker, M. (2013). Machine vs. human translation of SNOMED CT terms. Studies in health technology and informatics, 192, (pp. 581-584).

Kreuzthaler, M., Schulz, S., & Berghold, A. (2013) Top-Down Informationsextraktion aus klinischen Texten für die Sekundärnutzung der elektronischen Patientenakte.

eHealth2013 - Health Informatics meets eHealth - von der Wissenschaft zur Anwendung und zurück. (pp. 249-254).

Andrade, A. Q., **Kreuzthaler, M.**, Hastings, J., Krestyaninova, M., & Schulz, S. (2012). Requirements for semantic biobanks. In Proceedings of the 24th International Conference of the European Federation for Medical Informatics (MIE), 180(1), (pp. 569-573).

Kreuzthaler, M., & Schulz, S. (2012). Metonymies in medical terminologies. A SNOMED CT case study. In AMIA Annual Symposium Proceedings (Vol. 2012, pp. 463-467). American Medical Informatics Association.

Kreuzthaler, M., Bloice, M., Faulstich, L., Simonic, K. M., & Holzinger, A. (2011). A comparison of different retrieval strategies working on medical free texts. J. UCS, 17(7), (pp. 1109-1133).

Kreuzthaler, M., & Schulz, S. (2011, August). Truecasing clinical narratives. In Proceedings of the 23rd International Conference of the European Federation for Medical Informatics (MIE) - User Centred Networked Health Care (pp. 589-593).

Kreuzthaler, M., Bloice, MD., Faulstich, L., Simonic, K. M., & Holzinger, A. (2011, August) Mobile information retrieval in medicine: a semantic approach. In Proceedings of the 23rd International Conference of the European Federation for Medical Informatics (MIE) - User Centred Networked Health Care (pp. 272-274).

Bloice, M., Simonic, K. M., **Kreuzthaler, M.**, & Holzinger, A. (2011, November). Development of an interactive application for learning medical procedures and clinical decision making. In Proceedings of the 7th Conference on Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society: Information Quality in e-Health (pp. 211-224). Springer-Verlag.

Kreuzthaler, M., Bloice, M., Simonic, K. M., & Holzinger, A. (2011, November). Navigating through very large sets of medical records: an information retrieval evaluation architecture for non-standardized text. In Proceedings of the 7th Conference on Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society: Information Quality in e-Health (pp. 455-470). Springer-Verlag.

Bloice, M., **Kreuzthaler, M.**, Simonic, K. M., & Holzinger, A. (2010, November). On the paradigm shift of search on mobile devices: some remarks on user habits. In Proceedings of the 6th International Conference on HCI in Work and Learning, Life and Leisure: Workgroup Human-Computer Interaction and Usability Engineering (pp. 493-496). Springer-Verlag.

Kreuzthaler, M., Bloice, M. D., Simonic, K. M., & Holzinger, A. (2010, September). On the need for open source ground truths for medical information retrieval systems. In International Conference on Knowledge Management and Knowledge Technologies (i-KNOW) (Vol. 10, pp. 371-381).

Faulstich, LC., Müller, F., Sander, A., **Kreuzthaler, M.**, Kaiser, S., & Errath, M. (2009, May) Anwendung und Evaluierung Semantischer Retrievaltechnologien auf Medizinische Befundtexte. eHealth 2009 - Health Informatics meets eHealth - Von der Wissenschaft zur Anwendung und zurück. (pp. 41-47).